The University of Chicago

Where's My Bus: Using Automatic Vehicle Location Data to Identify Spatiotemporal Patterns of Bus Bunching in Chicago, IL

A thesis submitted in partial fulfillment of the requirements for a

BACHELOR OF ARTS IN PUBLIC POLICY STUDIES

at THE UNIVERSITY OF CHICAGO

Author: Annie Yang
Preceptors: Nitika Jain, Anthony Farmer

April 2020

Acknowledgements

Many times—every four hours or so—when I was writing this thesis, I wanted to drop the whole idea and start something new. The following people, places, and things kept me going: Daisy, Nancy, Krish, for their crash courses in stats and comp sci when the hour was late and the peril, great. Reynolds Club and Saieh, because who can actually do work in the Reg. The Ikea chairs in my apartment, for being so uncomfortable there was no way I would fall asleep during the one or two or five all-nighters pulled to finish this thing. My parents, for everything.

Abstract

This paper investigates the spatiotemporal distribution of bus bunching in Chicago. Using real-time bus location data pulled from the Chicago Transit Authority (CTA), this study conducts a thorough examination of two key features of bunching—when and where it tends to occur—with the urban geography of Chicago informing interpretations of what is found with respect to the research questions posed. The results of this study align well with what has been theorized about the bus bunching phenomenon from an engineering and operations research perspective. Bunching events fluctuate most in Chicago as a function of time to rush hour and distance from the city's central business district (CBD) in the Loop, which in turn corresponds with instances in which demand for service is high. These results inform recommendations regarding differ ways space can be prioritized for buses on Chicago's streets so that service reliability can be improved for public transit riders across the city.

Table of Contents

Introduction

Last summer, I joined some 200,000 commuters in their daily exodus to the Loop for work[1]. After work, I walked to the corner of State and Monroe and waited for the bus to take me back to Hyde Park. The minutes at the bus stop on State and Monroe were longer than usual—literally. More times than I can count, I would watch the prediction for my bus—arrival in 3 minutes, the overhead display would read—stay at 3 minutes for 7, then 10 minutes. As if to add insult to the injury, the bus, when it finally came, would often be accompanied by another bus that traversed the same route. This was a regularly infuriating experience that plagued my summer in Chicago.

The unreliability of Chicago's public bus system should certainly be cause for concern, not just because of the relatively minor inconvenience it posed for a college student on her commute to her summer internship. More broadly, public transit ridership in Chicago has declined at a year-to-year average of 3.1% since 2016. Chicago's transit agencies have cited various reasons for the decreased demand: cheap gas, popularization of ride-share, lack of funding[2][3]. Even as trends in public transit ridership shift, the importance of maintaining quality public transit in urban centers remains relevant, arguably more so than ever before. Public transportation is often the mobility lifeline for a city's low-income and disadvantaged populations[4]. Furthermore, widely implemented public transit is a more efficient method of minimizing congestion and environmental pollution in cities, relative to wide application of personal vehicle usage[5]. As such, while the effects of the variables described above (and others unmentioned) are nontrivial and deserve further study, it is worthwhile to consider what Chicago's public transit agencies *can* do to offer more competitive transportation options given the disruptions urban mobility has experienced in the last decade.

---

[1] "Where Workers Work - Labor Market Information." Illinois Department of Employment Security, November 2019. https://www2.illinois.gov/ides/lmi/pages/where_workers_work.aspx.
[2] Wisniewski, Mary. "Fewer Chicago-Area Residents Taking Buses, Trains." chicagotribune.com. Chicago Tribune, May 23, 2019.
[3] Clewlow, Regina, and Gouri Mishra. "Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States." *Journal of the Transportation Research Board*, October 2017.
[4] Criden , Madelaine . "The Stranded Poor: Recognizing the Importance of Public Transportation for Low-Income Households*." National Association for State Community Services Programs*, 2008
[5] Beaudoin, Justin, and C.-Y. Cynthia Lin Lawell. "The Effects of Urban Public Transit Investment on Traffic Congestion and Air Quality." *Urban Transport Systems*, 2017. https://doi.org/10.5772/66834.

In Chicago, the Chicago Transit Authority (CTA) manages bus and rail service for the city proper. On the CTA, most rides are taken on buses[6]. Within this geographic and operational context, I will investigate what measures can be taken to improve bus service on the CTA. In that regard, optimal bus systems in operations research has revealed that the phenomenon of bus bunching, where buses are unable to maintain their scheduled headway distances between one another and arrive at the same bus stop close to one another, frequently plagues bus systems across a wide variety of urban environments. In this paper, I investigate the spatiotemporal patterns of bus bunching in the CTA bus system. Doing so reveals which routes/bus stop would benefit the most from congestion relieving policies that aim to increase on time rates and improve overall performance. Findings from this research furthermore highlight the importance of evaluating the problem of bus bunching on a meso-scale, systemwide level, abstracting away from current literature's tendency to treat bus bunching on a route-by-route, case-by-case basis.

There are five sections in this paper. First, in the background section, I expound upon the introduction given here on the CTA, delving more in depth on the bus routes it offers and how route scheduling is managed. The background section also covers bus bunching as a phenomenon and the CTA's historical experience with it. Next, I present a literature review where bus bunching is the focus of the research to contextualize current academic approaches to addressing the issue. In the ensuing data and methods section, I begin with a description of the bus data used in the study and the methods used to analyze it. I then present the results of applying spatial analysis methods on the dataset. Finally, I conclude with some policy recommendations that reflect the findings presented previously.

[6] "About Us." CTA. Accessed December 2, 2019. https://www.transitchicago.com/about/

Background

In the previous section, I introduced bus bunching as an operational phenomenon and previewed its manifestation in urban transit systems. In this section, I expound upon these facets of bus bunching further by outlining the conceptual foundations of the bus bunching problem and discussing bus bunching's role as a reliability indicator in urban public transportation systems. I also provide an overview of the CTA bus system's performance to date in an effort to focus subsequent discussions of bus bunching to a uniquely Chicago context.

The Bus Bunching Problem. To adequately explain bus bunching, it is helpful to examine an ideal model of bus operation in an urban transit system. Consider a model with a single bus: service begins once the vehicle pulls out of a bus garage or yard and proceeds to a terminus stop on the route. From the terminus stop, the bus begins its route, altering between run time and dwell time states as it travels between successive stops (run time) to pick up passengers (dwell time)[7]. Inserting another bus into our single-vehicle example allows for another concept to be introduced: headway, a measurement of the space between two vehicles. Headway is typically calculated in terms of time (e.g., the buses are scheduled to maintain a headway of 5 minutes between each other), but can be calculated in terms of distance (e.g., the buses are scheduled to maintain a headway of 500 meters between each other) as well[8]. In this current model with no exogenous variables, the preservation of expected (scheduled) headway between the two buses is dependent on driver behavior and decision-making. As more variables (number of passengers at each bus stop, number of additional vehicles sharing the street with the buses, etc.) are added to the model, the factors impacting each bus's ability to maintain scheduled headway increases, thus decreasing the probability that scheduled headway can be maintained[9]. When buses are no longer able to maintain scheduled headway within a predefined margin of error, they are considered 'bunched.' The likelihood that bunching occurs also increases as more buses are added to a given route.

It is difficult to understate bunching's impact on service reliability and performance. When buses deviate off schedule—suppose for example, the first bus in the two-vehicle model above encounters traffic congestion and

[7] Ceder, Avishai. *Public Transit Planning and Operation: Theory, Modeling and Practice*. Amsterdam: Elsevier, 2007. p 28-29.
[8] Service Planning Terms and Concepts." King County Metro Transit. Accessed December 2, 2019. https://www.kingcounty.gov/transportation/~/media/transportation/kcdot/MetroTransit/HaveASay/Glossary.ashx
[9] GUENTHNER, RICHARD P. and KASIMIN HAMAT. "Distribution of Bus Transit On-Time Performance." Transportation Research Board, 1988. https://trid.trb.org/view/302165

arrives at its next stop late—dwell time at a bus stop increases as the bus must now wait longer for the build-up of passengers to board. Thus, in addition to the longer than expected wait times, the passenger load that the bus picks up is larger than expected[10]. The combination of long wait times at the bus stop and a ride experience on an over-capacity bus makes for an uncomfortable rider experience and reflects poorly on the bus system's performance.

Bunching and the CTA. Nationwide, headway adherence is something that most all transit agencies struggle with. An analysis from TransitCenter, a transit advocacy research organization, shows that among the twenty largest transit agencies across the country, the highest weekday on-time performance achieved by any agency was just 75% (on the Tri-county Metropolitan Transportation District of Oregon)[11]. In TransitCenter's study, the CTA ranks near the middle of the pack, with an on-time rate of 62%[12]. In 2012, the CTA received more complaints about bus bunching than any other grievance[13]. The trend continues into the present as bus bunching ranked number one on the top five complaints submitted to the twitter handle 'CTA Fails.'[14] The implications of unreliable service are significant for urban transit agencies. Study after study, both survey-based and empirical, confirms the fact that maintaining consistent, predictable bus service is essential for rider retention. The converse is also true: unreliable service as emerged as a commonality among transit systems that are experiencing declines in ridership[15,16,17,18].

To be sure, the CTA has taken measures to mitigate bunching on the 129 bus routes it operates in Chicago. On its website, the transit authority describes several preventative measures it takes to reduce schedule deviations on buses as well as mitigation techniques to minimize the effects of bus bunching when it occurs. Preventative

---

[10] Bowman, Larry A., and Mark A. Turnquist. "Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops." *Transportation Research Part A: General* 15, no. 6 (1981): 465–71. https://doi.org/10.1016/0191-2607(81)90114-x.
[11] "Your Bus Is On Time. What Does That Even Mean?" TransitCenter, August 13, 2019. https://transitcenter.org/bus-time-even-mean/.
[12] *Ibid*
[13] Bartholdi, John J., and Donald D. Eisenstein. "A Self-Coördinating Bus Route to Resist Bus Bunching." Transportation Research Part B: Methodological 46, no. 4 (2012): 481–91. https://doi.org/10.1016/j.trb.2011.11.001.
[14] "Top 5 CTA Complaints via CTA Fails - Chicago Tribune." Chicago Tribune, February 9, 2016. https://www.chicagotribune.com/redeye/redeye-top-5-cta-fails-complaints-20160120-story.html.
[15] Perk, Victoria. "Transit Ridership, Reliability, and Retention." *National Center For Transit Research*, January 2008. https://doi.org/10.5038/cutr-nctr-rr-2005-08.
[16] Higashide, Steven, and Mary Buchanan. "Who's on Board." TransitCenter. Accessed December 2, 2019. https://transitcenter.org/publication/whos-on-board-2019/.
[17] Currie, Graham, and Ian Wallis. "Effective Ways to Grow Urban Bus Markets – a Synthesis of Evidence." *Journal of Transport Geography* 16, no. 6 (November 2008): 419–29. https://doi.org/10.1016/j.jtrangeo.2008.04.007.
[18] Bates, John, John Polak, Peter Jones, and Andrew Cook. "The Valuation of Reliability for Personal Travel." *Transportation Research Part E: Logistics and Transportation Review* 37, no. 2-3 (April 2001): 191–229. https://doi.org/10.1016/s1366-5545(00)00011-9.

measures include building in additional 'slack' time between buses to account for potential delays they may encounter while running a route. In some cases, the CTA will actually opt to increase the number of buses running a route (even though conventional knowledge suggests that bunching is more common on high-frequency routes) in the event that it anticipates delays to be compounded by crowding[19]. Since 2009, the CTA has rolled out GPS systems on its buses to provide real-time information about bus location in an effort to help both itself and its passengers track bus [20]. In instances where buses have deviated so far from their schedule so much so that bunching is imminent, the CTA implements delay mitigation measures like running incoming buses express and rerouting a bus back to its terminus point before it completes a full iteration of its route[21].

Transit agencies like the CTA must consider myriad factors when devising an optimal bus schedule. On the supply side, it must predict the optimal number of buses to operate on each route throughout each service period during the day while remaining cognizant of the tradeoff between service frequency and reliability. How much service the agency ultimately decides to provide must most importantly match passenger demand. Thinking further about demand, so much of the decision rests on ensuring that service matches passenger expectations. In this, passenger expectation is not so much that a bus arrives at its stop at a certain time, but that any bus on a given route should arrive at its stop on regular intervals. The expectation of riders is that they should be able to get to a bus stop and, in not too long time, board a bus. Because so much of the problem at hand is about matching these rider expectations, it is useful to know where and when bus bunching tends to occur. The CTA, equipped with knowledge not only about which routes tend to experience bunching, but which areas in the city experience bunching most frequently, would be able to hedge against delays across its whole system more efficiently and thus achieve greater returns on whatever combination of strategies they use to recuperate lost headway adherence. In the subsequent literature review, I show how current literature, while helpful to transit agencies for managing bus scheduling on a

[19] "When Things Go Wrong (How We Mitigate Delays)." CTA. Accessed December 2, 2019. https://www.transitchicago.com/performance/wtgw/#bussystem.
[20] "New CTA bus tracker service alert feature arriving in addition to more routes on March 2." CTA. February 27, 2009. Accessed December 2, 2019. https://www.transitchicago.com/new-cta-bus-tracker-service-alert-feature-arriving-in-addition-to-more-routes-on-march-2/
[21] ''When Things Go Wrong (How We Mitigate Delays)"

route-by-route basis, has historically fails to consider these system wide interactions between routes. My research seeks to close this gap by treating bus bunching as foremost a problem of space.

Literature Review

Research for this thesis is conducted under the assumption that there is value in addressing the problem of bus bunching for public transportation agencies operating in urban areas. This premise is supported by the wealth of prior research on the importance of service reliability as it pertains to customer retention for public transit systems. As early as the 1960s, transit researchers have investigated how individuals decide what mode of transportation to take when they travel [22]. The attitudinal approach of these studies indicated from early that the reliability of a transportation option was of great concern to riders surveyed[23, 24]. Findings from more recent research continue in this vein. In the well-cited study 'The Valuation of Reliability for Personal Travel,' Bates et al. find that 'punctuality is indeed highly valued by travelers.' For public transportation in particular, the authors note there is also disutility for riders in mistiming travel and arriving too earlier and/or late at a destination. Later quantitative analyses of the relationship between ridership and service reliability reaffirm the findings from attitudinal surveys of years past. Among these, one study by Perk et al. finds strong positive correlations between service quality (a composite of different variables including 'reasonable wait time,' 'fast travel time,' 'smooth, uncrowded ride,' among others) and customer satisfaction[25]. Perk's research suggests that bunching is not only a problem for transit agencies because of the real drops in service quality it engenders, but also because of the perception of unreliability that it can generate that additionally damages the image of the transit agency and decreases subsequent ability to sustain demand for public transportation services.

On the topic of bus service reliability, it also useful to review non-academic literature published by independent research organizations and advocacy groups that frequently collaborate with public transit agencies to evaluate transit policy. In these instances, bus bunching is frequently grouped under a larger umbrella of transit reliability/performance. A report released by TransitCenter, a leading transit advocacy and research group, refers to bunching in its analysis of service unreliability on the Metropolitan Transportation Authority buses in New York

[22] Hill, S. J., and G. A. Brunner. "Consumer Conceived Attributes of Transportation." University of Maryland, 1967.
[23] Bates, John, John Polak, Peter Jones, and Andrew Cook. "The Valuation of Reliability for Personal Travel." *Transportation Research Part E: Logistics and Transportation Review* 37, no. 2-3 (2001): 191–229. https://doi.org/10.1016/s1366-5545(00)00011-9.
[24] Prashker, Joseph N. "Direct Analysis of the Perceived Importance of Attributes of Reliability of Travel Modes in Urban Travel." *Transportation* 8, no. 4 (1979): 329–46. https://doi.org/10.1007/bf00167987.
[25] Perk, Victoria. "Transit Ridership, Reliability, and Retention ." National Center for Transit Research, 2008.

City[26]. Another advocacy group, Active Transportation Alliance, conducted a survey of over two thousand riders on the CTA and found that 'the most frequent suggestion from riders we surveyed for improving service was to eliminate bus bunching,' further highlighting why its important for transit agencies to address bunching; the phenomenon is easy to identify and riders immediately associate its occurrence with poor service[27].

It becomes clear that the bus bunching problem deserves a solution when taking into considering the corpus of transit reliability research as a whole. In this, academic research largely treats bus bunching as an operations problem. Leveraging concepts from civil and industrial engineering, researchers have proposed myriad methods to prevent and reduce bunching on buses across a whole host of urban transit systems. Common practices for addressing the problem of bus bunching typically call for some combination of slack and holding to mitigate the snowball effect that takes place when bunching occurs. In this, bus bunching mitigation is often treated as an optimization problem; a common objective being to estimate optimal amounts of slack and hold time in an effort to help transit agencies make the most efficient tradeoff between service frequency and reliability[28 29 30]. Until recently, many studies relied on simulating the behavior of buses to study bus bunching, resulting in findings that are heavily dependent on the assumptions made by the authors of the study[31,32,33]. In the last decade, automatic vehicle location (AVL) tracking has become more common as GPS tracking devices have been installed on buses in many of the nation's largest public transit systems. Researchers have taken advantage of the new data, incorporating both AVL

---

[26] "Turnaround: Fixing New York City's Buses - TransitCenter." TransitCenter. Accessed December 10, 2019. http://transitcenter.org/wp-content/uploads/2016/07/Turnaround_Fixing-NYCs-Buses-20July2016.pdf.

[27] "BACK ON THE BUS: SPEEDING UP CHICAGO'S BUSES." Active Transportation Alliance. Accessed December 10, 2019. http://activetrans.org/sites/files/SpeedingupBuses.pdf.

[28] Lee, Kurt K. T., and Paul Schonfeld. "Optimal Slack Time for Timed Transfers at a Transit Terminal." *Journal of Advanced Transportation* 25, no. 3 (1991): 281–308. https://doi.org/10.1002/atr.5670250304.

[29] Zhao, Jiamin, Maged Dessouky, and Satish Bukkapatnam. "Optimal Slack Time for Schedule-Based Transit Operations." *Transportation Science* 40, no. 4 (2006): 529–39. https://doi.org/10.1287/trsc.1060.0170.

[30] Hall, Randolph, Maged Dessouky, and Quan Lu. "Optimal Holding Times at Transfer Stations." *Computers & Industrial Engineering* 40, no. 4 (2001): 379–97. https://doi.org/10.1016/s0360-8352(01)00039-0.

[31] Seneviratne, Prianka N. "Simulation of Fixed Route Bus Travel Time." Journal of Advanced Transportation 22, no. 1 (1988): 39–53. https://doi.org/10.1002/atr.5670220104.

[32] Powell, Warren B., and Yosef Sheffi. "A Probabilistic Model of Bus Route Performance." Transportation Science 17, no. 4 (1983): 376–404. https://doi.org/10.1287/trsc.17.4.376.

[33] Chapman, R. A., and J. F. Michel. "Modelling the Tendency of Buses to Form Pairs." *Transportation Science* 12, no. 2 (1978): 165–75. https://doi.org/10.1287/trsc.12.2.165.

data and automatic passenger count (APC) data into their models and simulations[34,35]. New recommendations based on more readily accessible AVL and APC data have called for bus hold times to be calculated on the fly in order to reduce the amount of slack needed to be embedded into bus schedules[36,37,38]. Beyond bus bunching as an operations problem, some non-operational research on bus bunching does exist, but mostly focuses on the performance of one or two routes within a larger system. In this space, the body of regression-based research on this topic is expanding. It has provided consensus on variables like passenger demand, on-street traffic conditions, behavior of bus operators, route features, and proximity to route termini as factors associated with the occurrence of bus bunching along a route[39,40,41]. Where the literature varies is in its treatment of how the bus bunching problem should be modelled. Many techniques, stemming from a wide variety of assumptions about the form of the trends experienced by bus bunching have been applied, from simple logistic regressions, to transformed autoregressive ones, to non-parametric treatments of the problem[42,43,44,45].

Reflecting on existing literature, it is my goal in this thesis to largely continue in the tradition of existing research, both academic and nonacademic. However, my paper will place greater weight on studying the spatial-temporal dimensions of bunching, which has been featured less prominently in the survey of literature conducted for this review. Furthermore, I take the problem of bus bunching as given, focusing more then, on answering the

---

[34] Andres, Matthias, and Rahul Nair. "A Predictive-Control Framework to Address Bus Bunching." *Transportation Research Part B: Methodological* 104 (2017): 123–48. https://doi.org/10.1016/j.trb.2017.06.013.
[35] Pi, Xidong, et al. "Understanding Transit System Performance Using AVL-APC Data: An Analytics Platform with Case Studies for the Pittsburgh Region." *Journal of Public Transportation* 21, no. 2 (2018): 19–40. https://doi.org/10.5038/2375-0901.21.2.2.
[36] Daganzo, Carlos F. "A Headway-Based Approach to Eliminate Bus Bunching: Systematic Analysis and Comparisons." *Transportation Research Part B: Methodological* 43, no. 10 (2009): 913–21. https://doi.org/10.1016/j.trb.2009.04.002.
[37] Bartholdi, John J., and Donald D. Eisenstein. "A Self-Coördinating Bus Route to Resist Bus Bunching." *Transportation Research Part B: Methodological* 46, no. 4 (2012): 481–91. https://doi.org/10.1016/j.trb.2011.11.001.
[38] Pilachowski, Joshua Michael. "An Approach to Reduce Bus Bunching." University of California, Berkeley, 2009.
[39] Rashidi, Soroush et al. "Using Automatic Vehicle Location Data to Model and Identify Determinants of Bus Bunching." *Transportation Research Procedia* 25 (2017): 1444–56. https://doi.org/10.1016/j.trpro.2017.05.170.
[40] Yu, Haiyang, et al. "Probabilistic Prediction of Bus Headway Using Relevance Vector Machine Regression." *IEEE Transactions on Intelligent Transportation Systems* 18, no. 7 (November 15, 2016): 1772–81. https://doi.org/10.1109/tits.2016.2620483.
[41] Jiang, Rui-Sen, Da-Wei Hu, and Xue Wu. "Prediction of Bus Bunching Using Smart Card Data." *Cictp 2019*, February 2019. https://doi.org/10.1061/9780784482292.122.

[42] Verbich, David, Ehab Diab, and Ahmed El-Geneidy. "Have They Bunched Yet? An Exploratory Study of the Impacts of Bus Bunching on Dwell and Running Times." *Public Transport* 8, no. 2 (2016): 225–42. https://doi.org/10.1007/s12469-016-0126-y.
[43] Iliopoulou, Christina, et al. "The Bus Bunching Problem: Empirical Findings from Spatial Analytics." *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018. https://doi.org/10.1109/itsc.2018.8569760.
[44] Bharathi, B. Dhivya et al. "Bus Travel Time Prediction: a Log-Normal Auto-Regressive (AR) Modelling Approach." *Transportmetrica A: Transport Science* 16, no. 3 (January 2020): 807–39. https://doi.org/10.1080/23249935.2020.1720864.
[45] Rashidi, Soroush et al. "Using Automatic Vehicle Location Data to Model and Identify Determinants of Bus Bunching."

question: where and when does bus bunching occur in the transit context of Chicago, and most crucially, why does bus bunching happen during those times, at those places? Regarding this objective, to the best of the author's knowledge, there has been no prior attempt to produce stop-level predictions of where bus bunching is likely to occur in a given study area. In that the goal is to better urban transportation policy, this paper's primary objective is to more fully scope the circumstances in which bus bunching occurs in Chicago. Treating bunching foremost as a problem of space, I seek to better depict the context that bus bunching both experiences and creates so that the operational improvements proposed in academia and advocacy may be applied in the most effective manner, given Chicago's unique urban characteristics.

Methods and Data

In this section, I describe the data and methods used to determine where and when bus bunching occurs in Chicago. I also present an approach for predicting where bunching in the network is likely to occur in the future.

Methods. The methods used to answer the questions of where, when, and why bus bunching occurs in Chicago can be broken into two parts. Overarching them is the question of what scale further analyses should take place on, which is considered first here. On this matter, the literature review conducted for this thesis suggests that much of the existing research investigates bunching at the route level. While such an approach has its advantages, it neglects both more minute observations at the stop level and between-route interactions at the network level. As such, there is a deliberate attempt in this thesis to apply the methods described below at all three scales of analysis—network, route, and stop.

The first part of this study's methodology is relatively straightforward—it involves measures for tracking bus bunching—where and when it occurs—in the present. For the latter, the temporal distribution of bunching incidents can be explored using histograms, as well as other depictions of spread. With the former, data gathered on the location of bus bunching is used to answer the question of where bus bunching occurs. To visualize where buses are bunched, I apply techniques for creating heatmaps in QGIS and Python to map the spatial distribution of buses, per route, over the time it takes for the bus to run its route. Aggregating individual bus route heatmaps over longer periods of time can then produce a heatmap of where buses tend to bunch across the entire CTA bus system. These techniques, while similar in concept, vary in form at different scales of analysis. For example, to understand the spatial distribution of bunching at the network level, a kernel density estimate of incidents is calculated. At the route level, streets are discretized into 500-meter segments, and incidents are grouped to each segment accordingly. Moreover, the directionality of routes is also taken into account at the route level; it is however, left aggregated when considering the bunching patterns at the network level, as segregating by direction of travel would omit import interactions between routes when two or more intersect. In thinking about how route segments should be clustered or which bus stops have high/medium/low levels of bunching, the most appropriate measure that emerged was Jenks natural breaks. This classification approach minimizes within group variance and maximizes intergroup variance, which fits the objective of this analysis.

The second part of my analysis takes more effort to structure. To reiterate, I construct a model to predict where future instances of bus bunching will occur. First, it may be helpful to clarify why such an analysis is necessary. Simply put, more insight than where and when bus bunching occurs is needed to make a meaningful recommendation of how public transit can be improved in Chicago. From a policy perspective, identifying common attributes between bus stops that experience bus bunching more directly helps transit agencies formulate better strategic planning for improved future service.

Now that motive is clarified, let us now consider approach. In that I want to estimate the likelihood that bunching occurs at bus stop *b* given some attributes associated with the bus stop at time *t*, conducting a Poisson regression emerges as an obvious path forward, since the response variable of interest here is a count of bunching incidents grouped by bus stops. Indeed, many studies that simulate bus travel times and demand assume that buses arrival time at stops follows a Poisson distribution[46,47,48.]. However, Poisson regressions require mean-variance equality in the distribution of data. Since this characteristic is difficult to fulfill with most any real-world data, it will be useful to compare the fits of the Poisson regression with the results of other transformations, such as a negative binomial (NB) regression. The NB regression is commonly used when count data is over-dispersed (i.e. when mean and variance are not equal), as is likely the case here. Another adjustment that may be worthwhile to consider is using a zero-inflated version of either model in the event that there are an excessive number, that is, more than one would expect in a Poisson or NB distribution, of bus stops with zero bunching incidents.

The response variable for whichever model is ultimately selected is easy to define, simply the number of bunching incidents per bus stop, the predictor variables are more difficult to select. The review of literature

---

[46] Zhao, Shuzhi, et al. "A Self-Adjusting Method to Resist Bus Bunching Based on Boarding Limits." *Mathematical Problems in Engineering* 2016 (2016): 1–7. https://doi.org/10.1155/2016/8950209.

[47] Koppisetti, M Venkateswararao et al. "Bus Schedule for Optimal Bus Bunching and Waiting Times." *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, 2018. https://doi.org/10.1109/comsnets.2018.8328282.

[48] Saw, Vee-Liem, and Lock Yue Chew. "No-Boarding Buses: Synchronisation for Efficiency." *Plos One* 15, no. 3 (2020). https://doi.org/10.1371/journal.pone.0230377.

conducted for this study suggests that a key generative factor of bus bunching is passenger demand. Variability in the time that buses dwell at stops along a route invariably underlie many of the variables that have been identified in previous studies as closely related to bus bunching occurrences. On popular bus routes, more buses are scheduled to accommodate for higher demand, inevitably increasing the probability of bunching. To capture this, a variable is introduced in the model to capture the effect of more frequent bus service, which, at its base, a function of higher passenger demand. This should help return more accurate estimates of the effect of the independent variables ultimately included the model.

Another important factor to account for in constructing the right-hand side of this NB model is the spatiotemporal nature of the data being used. One concern in this dimension is the level of bunching faced during rush hour, which may lead to biased estimates if left uncontrolled. Furthermore, this feature of the data invites autocorrelation; intuitively the occurrence of bus bunching at one stop affects whether bunching will occur at subsequent stops. Finding the best way to account for autocorrelation requires testing, but common corrections include adjusting the functional form of the model, adding/deleting variables, and increasing the interval between observations (e.g. estimating outcomes for every tenth bus stop in a route, rather than each bus stop). In the context of this study, it seems sufficient to simply include a binary variable that notes whether bunching has occurred at the previous stop for a given stop as a sort of modified 'lag' term to account for the effect of autocorrelation. After additionally controlling for socioeconomic effects, the base model takes the form:

$$Counts = \beta_0 + \beta_1 Demand + \beta_2 Demand * Rush\ Hour + \beta_3 Rush\ Hour + \beta_4 Route\ Features + \beta_5\ Socioeconomic\ Attributes\ + \beta_6 Bunching\ at\ previous\ stop +\ \varepsilon$$

Where socioeconomic variable and route feature variable both represent a vector of terms. For variables that are not inherent attributes of bus stops (e.g., a stop has no median household income, per se), values from the closest Census block group or Census tract in which a bus stop is located will be used. Testing the final model would involve assessing its probability estimates for out of sample data, which can be arranged by splitting the bus bunching data collected in the first part of this analysis into train/test datasets.

To end this methods section, I note some limitations to the approach specified above and highlight some alternatives that are still being considered at this point in the study process. By far, the largest concern with the analysis as described is poor specification of the model due to omitted variables, endogeneity, and other factors unforeseen. A solution, not to be explored in this study, to the difficulty that constructing a well specified regression presents in this research context is to simulate bus bunching behavior based on known attributes of where bunching occurs throughout the CTA bus system historically.

Data. This study requires a minimum of two types of bus data: location and time. Data on both attributes can be collected in real-time using the CTA's BusTracker API. In particular, the API's *getvehicles* and *getpredictions* functions constitute the main avenues by bus location and time data are collected. Each function has certain parameters that are of interest to my analysis. In *getvehicles*, the 'tmstmp,' 'lat,' and 'lon' parameters represent the time and location values for a bus per *getvehicles* call. The latter two parameters are necessary for capturing the location of buses while the former parameter reports the time at location. Together, these parameters are essential for determining whether there are buses clustered together at a given time in a way that suggests that bus bunching is occurring. In *getpredictions,* the 'dstp,' 'prdtm', 'dly', and 'prdctdn' parameters provide information on a bus's distance to the next stop, its predicted time of arrival, delay status, and time left until arrival, respectively. These variables offer greater detail into how the CTA schedules buses and will be invaluable for classifying bus routes by their on-time performance.

Furthermore, on the topic of data collection, it is desirable to maximize the amount of bus location data collected, both in terms of duration and in terms of bus system completeness. Ideally, data for all bus routes (129 routes) run by the CTA, can be recorded. Given the time constraints of this study, I have collected 280 hours of data for seventeen bus routes across the CTA service area totaling in 1,805,296 unique observations. These bus routes are a part the CTA's Night Owl Service, which is a "core network of bus and rail service that operates 24 hours per day, seven days per week." [49] In the CTA's own words, the Night Owl buses are a representative subset

[49] "Night Owl Service." *Chicago Transit Authority*, 2018. https://www.transitchicago.com/assets/1/6/ctamap_OwlService.pdf.

of the entire CTA bus system, both in terms of coverage area and in terms of ridership makeup. Based on most recently available data, nine out of the CTA's ten highest ridership routes are Night Owl routes. As such, they will do well in this study as a proxy for the whole network, providing the necessary data for modeling transit flow systemwide.

Bus Number	**Route Name**	**Ridership Rank as of Jun 2019**	**Direction**	**Route Information**
4	Cottage Grove	3	N-S	Runs four different routes between 115th/Cottage Grove and Columbus/S. Water. Overnight service runs between 63rd/Cottage Grove and Columbus/South Water
6	Jackson Park Express	32	N-S	Runs between 79/South Shore to Wacker/Columbus. Overnight service covers South Shore area between 69th and 95th streets (this study considers the 5 South Shore Night Bus to be the Night Owl Service of the 6 Jackson Park Express)
9	Ashland	5	N-S	Runs four different routes between Ashland/95th and Clarke/Belle Plaine. Overnight services runs between 95th/Dan Ryan and North/Clark
20	Madison	12	E-W	Runs between Madison/Austin to Randolph/Columbus
22	Clark	8	N-S	Runs between Howard to Clark/Harrison
34	South Michigan	60	N-S	Runs between 131st/Ellis and 95th/Dan Ryan
49	Western	9	N-S	Runs on Western Ave between 79th and Berwyn
53	Pulaski	7	N-S	Runs on Pulaski between Harrison and Irving Park. Night service between Pulaski/31st and Pulaski/Peterson
55	Garfield	34	E-W	Runs on Garfield Blvd between Midway and Museum of Science and Industry. Night service between MSI and St. Louis
60	Blue Island/26th	31	E-W	Runs between Cicerp/24th Pl and Harbor Dr/Randolph. Night service between Washington/State and 54th/Cermak

62	Archer	28	N-S	Runs between Archer/Harlem and State/Kinzie. Night service between State/Washington and Midway
63	63rd	14	E-W	Runs on 63rd St between Midway and 63rd/Stony Island
66	Chicago	2	E-W	Runs between Chicago/Austin and Navy Pier. Night service between State/Washington and Chicago/Pulaski
77	Belmont	4	E-W	Runs four routes between Belmont/Cumberland and Lake Shore/Diversey. Night service between Halsted and Harlem
79	79th	1	E-W	Runs on 79th St from Western to Lakefront.
81	Lawrence	26	E-W	Runs on Lawrence between Jefferson Park and Wilson/Marine
87	87th	25	E-W	Runs four routes 88th/Cicero and 91st/Commercial. Night service between Dan Ryan and Western on 87th St

Table 1. CTA Owl Bus Route descriptions

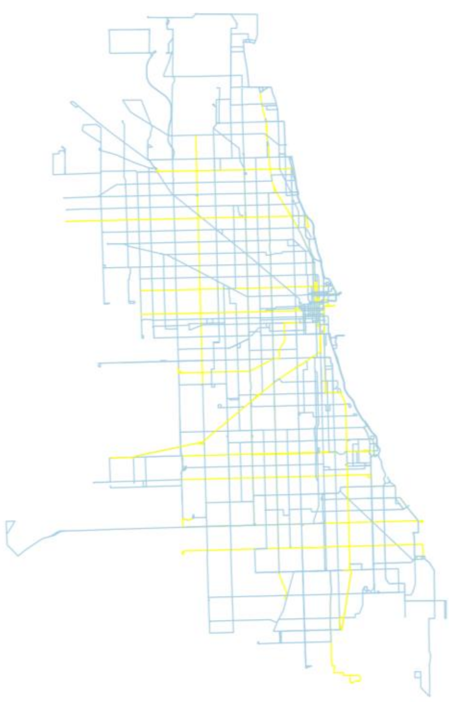

Fig 1. Map of CTA bus routes. Night Owl routes are highlighted in yellow.

The information collected from the BusTracker API constitutes the bulk of the data used in this study. Most of the variables in the NB model can be constructed from analysis of the primary data sources themselves. Attributes that are part of the 'Route feature' matrix described in the methods section—whether bunching occurred during rush hour, bunching presentation at adjacent stops, direction of the route stops served—can most all be derived from the time and location data already collected. Attributes in this collection that cannot be extracted from the primary source, of which there is only one, stop proximity to places of interest (i.e. employment centers; it is imaginable that popular destinations of travel are also where many people work), is identified using the latest Census employment density data. To control for socioeconomic effects, median household income at the block group level is assigned to bus stops that fall within that given Census block. Beyond these sources, a few more supplements are necessary.

As discussed previously, passenger demand is a key variable of interest. While the CTA does possess the capability to collect boarding counts per stop, I was unable to obtain this data even after multiple attempts to contact the CTA for the relevant information. As such, a proxy for passenger demand must be implemented instead. Finding a suitable substitute for passenger demand is challenging to stay the least, since the factors that influence how many boardings a stop experiences are manifold and varying. To this end, two sources will be consulted, population density at the Census block group level from the American Community Survey, and a publicly available CTA dataset of passenger counts by bus stop for the month of October 2012. Using these sources is obviously not ideal, but the decision to include them extend from generally reasonable assumptions: bus stops located in more populous areas should be frequented more, and whatever the factors that influence where passengers board are, they should be mostly stationary over time; in the short term at least, stops that are historically popular should not experience drastic changes in status. The table below summarizes the data backing each of the predictor variables considered for addition to submission count regression models.

Variable	Attribute	Source
Demand	Pop density	American Community Survey five-year estimates, 2014-2018
Demand	Historical passenger count by bus stop	Avg. Weekday Bus Stop Boardings in October 2012
Rush hour	Incidents occurring during rush hour/non-rush hor periods	AVL data collected from BusTracker API
Lag term	Bunching at previous stop	AVL data collected from BusTracker API
Socioeconomic	Mode of transportation to work	American Community Survey five-year estimates, 2014-2018
Socioeconomic	Median Household Income	American Community Survey five-year estimates, 2014-2018
Route feature	Stop direction of travel	CTA Bus stop shapefile
Route feature	Position on street	CTA Bus stop shapefile
Route feature	Services places of interest	2017 Census LEHD employment density data
Route feature	Number of routes served	CTA Bus stop shapefile

Table 2. Data sources for explanatory variables

Analysis

In this section, I describe how the methods laid out in the previous section were applied on real-time bus data that was collected over two weeks in early January. At a high level, the goal was to visualize and relate the spatial temporal manifestations of bus bunching at various degrees of aggregation. Bus data sourced from multiple bus routes allows us to study the space-time dynamics of bus bunching at three levels: stop, route, and network. First though, I introduce the data used in more detail and discuss was learned in a first past through the data. I also explain what constituted bus bunching in this analysis, and the rationale for defining bus bunching in such ways.

Data Description. Over a two-week period of data collection, AWS Lambda functions pulled in 1,805,296 unique bus locations across the seventeen routes previously specified through the BusTracker API. Bus location data came enclosed in a JSON file, which was then flattened and parsed into a data table for further analysis. A sample of the master data table is included below for ease of reference:

des	dly	hdg	lat	lon	pdist	pid	rt	tablockid	tatripid	tmstmp	vid
Kinzie/State	False	50	41.832727	-87.673825	41901	7110	62	62 -613	106056	2020-01-07 00:02:10	8018
Kinzie/State	False	267	41.800507	-87.725723	22023	7111	62	62 -656	340	2020-01-07 00:02:12	1521
Kinzie/State	False	358	41.866917	-87.627388	61201	7111	62	N62 -692	41	2020-01-07 00:02:01	8093
Kinzie/State	False	6	41.881877	-87.629333	67037	7111	62	N62 -694	40	2020-01-07 00:02:05	8220
Harlem	False	235	41.853080	-87.634836	14436	7120	62	N62 -691	106177	2020-01-07 00:01:52	8000

Table 3. Raw BusTracker API data, in tabular form

Each row in the data table represents a vehicle at a given time and location, as noted by the 'vid,' 'tmstmp,' and 'coord' columns respectively. Along with information about the location of a bus at a given time, the API call also returned relevant information regarding the direction of travel, delay status, cardinal direction of the bus, and distance remaining (in feet) in the current route being executed. The first row, for example, can be interpreted as:

“Bus 8018 is traveling towards Kinzie/State on the 62 Archer route at 2:10 AM. Its current coordinates are (-87.674, 41.832).”

Valuable insights can be derived from just the raw data itself. Some simple filtering and pivoting reveal precursory, but valuable nonetheless, information on route demand and performance. By grouping each row, representing a vehicle, by day and route, it is possible to find the average number of buses that run a given route each day. Pairing this information with the total number of delays each route experiences over the entire study period produces a somewhat intuitive relationship between route demand and number of delays. The 49 Western route in particular, in addition to having the highest average buses running each day over the study period, also experienced the most delays in service over the same time. The implication that offering more regular service comes with a tradeoff of less consistent service is certainly of interest from the policymaker’s perspective in transportation planning and deserves further investigation in the analyses to follow.

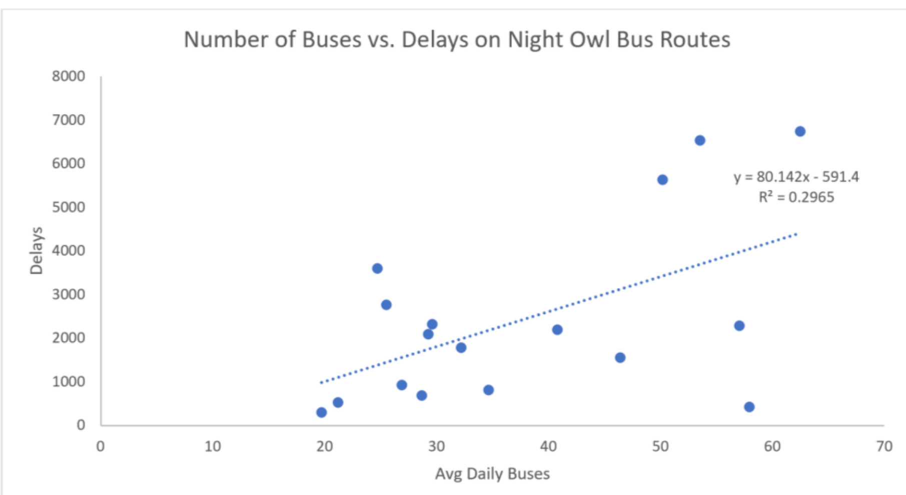

Fig 2. Number of Buses on a Route vs. Delays Experienced by Route

Bus Bunching Identification. Before proceeding, some clarification on how bus bunching is defined in this study is needed. As previously stated, the spacing between two buses can be measured in distance or time. For example, buses can have a scheduled headway of ten minutes, that is to say, if a bus arrives at a stop at time t, the next bus should arrive at the station $t+9$ minutes after the previous bus. According to the CTA’s own statistics, 87%

of all CTA bus customers ride routes where the headway is 12 minutes or less[50]. In the worst cases of bunching, there will be cases in which the headway between two buses is 0 minutes: the following bus is right behind, or overlapping with, the preceding bus. When such severe instances of bunching occur, the distance between two buses is effectively zero. That said, identifying what the upper bound for a distance between buses should be for the two buses to be considered bunched requires slightly more effort to determine. For this task, the average distance between two consecutive buses at a given time, for all intervals in the study was calculated. Accounting for these factors along with the fact that the most severe cases of bus bunching are what is of interest for this study, a minimum threshold of .003 km (~10 ft.) between buses was chosen as the distance between buses smaller than which would be considered a bunching incident. After using this threshold to identify cases of bus bunching, the midpoint between the two coordinate locations of the buses was calculated to demarcate the incident.

For another perspective on bus bunching, positions of buses relative to bus stops were also considered. In this approach, for a given route, direction, and time, bus coordinates are plotted on top of bus stops along the route. Each bus coordinate is assigned a bus stop based on distance, that is under the given parameters, buses are assigned to the bus stop that is currently closest to them. For a set of parameters, a bus stop is said to be experiencing bunching when two or more busses are assigned to the same stop. The diagram below models this form of bunching identification. The buses, represented by rectangles, are distributed along an eastbound traveling route. The stop in the middle, has two buses that are closer to it than any other stop. Hence, the bus stop encountered bus bunching and is highlighted in red. In this way, the analysis conducted for this study is able to capture how bus bunching is distributed across space and time at not only the route and network levels, but at the stop level as well.

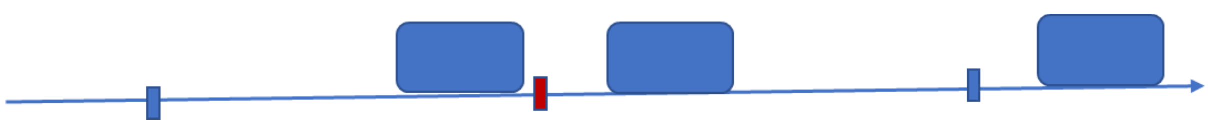

Fig 3. Model of stop-based bunching identification.

[50] "Performance Indicators." *Chicago Transit Authority,* https://www.transitchicago.com/assets/1/6/0602opperf.pdf

An important consideration accounted for during the identification process was devising a way to differentiate between true bunches and scheduled waiting periods in which bus congregation is expected behavior. At the end of routes, for example, buses often wait a designated period of time before resuming their route. Both approaches to bunching identification used in this study rely on some combination of time and distance variables to classify whether buses are bunched. As bunching identification currently stands, there is room for a nontrivial number of false positives, that is, classifying two or more buses as bunched when in fact their proximity to one another is expected behavior. An example of this is when capacity is intentionally increased during high demand hours. To address this issue, a possible countermeasure is to down weight bunchings that are close to the terminal stops of a route.

Bunching at the Network Level. Across the CTA bus network, as proxied by the seventeen core routes studied, 8,285 bunching incidents were identified using the distance-between method with a threshold of ten feet in the two-week data collection period between Jan 8 and Jan 22. It should be noted that incidents are recorded every two minutes, meaning that any given instance of bunching between two of the same bunches is recorded until the end of the bunching event. The rationale for doing so despite there being an overestimate in bunching events is that recording incidents at such a granular time interval allows data to be collected on the duration of bunching incidents, in addition to where they tend to occur.

Plotting the entire set of data reveals a map with incidents so frequent that they approximate the rotes on which they run (Fig 5). Without any further form of manipulation or calculation, the concentration of bunching incidents that occur in the Downtown Loop area is already visible. Applying kernel density estimation (KDE) to the data further reinforces this observation (Fig 6). Outside the Loop, the heatmap created using KDE indicates concentrations of incidents at the terminuses of routes. Recalling that the likelihood of false positive incidents being higher at the ends of stops, a closer investigation of the particularly high concentrations of incidents on the north side and south side reveal that the collection of points take place around the Western and 79th Bus terminuses respectively. These incidents should be weighted appropriately when interpreting the implications of this network level analysis. Furthermore, darker points are visible at various intersections where different routes cross. To investigate the relationship between bunching incidents and route intersections, the spatial extent of the network was discretized into grids of equal size and the number of points in each grid was calculated (Fig 4). Grids that contained a point of intersection between routes were subsequently identified and labelled as such. When the intersection of two or more routes fell between grids, any grid with an edge touching the intersection was assigned as an intersection grid. Removing empty grids from the calculation, on average, every grid contained 29.48 incidents. Intersection grids, meanwhile, contained an average of 41.94 incidents. Although it is unclear whether the higher number of incidents around intersections is statistically significant, intuitively, it makes sense that points along routes that intersect with other routes would be subject to greater variance in on-route events, thus increasing the number of variables that can disrupt a bus's ability to adhere to its schedule.

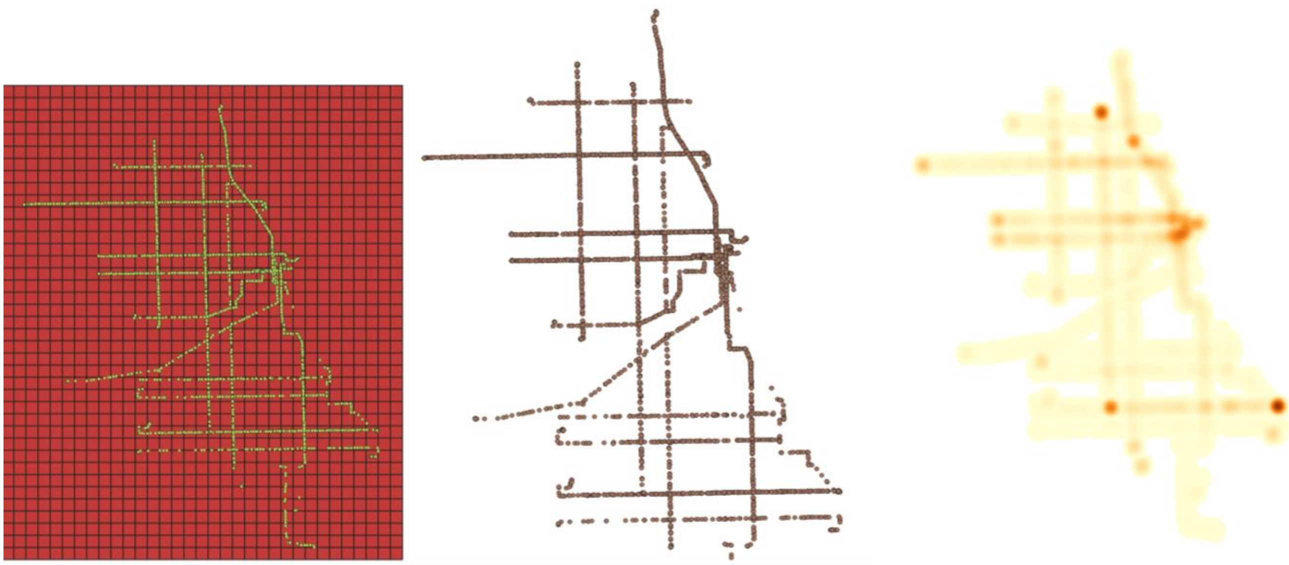

Fig 4, 5, 6. Discretized plot of bunching events. Point and KDE plots of bus bunching at the network level

To incorporate time variables into analyses, the set of bunching incidents was cut by hour and day. Doing so generated insight into how levels of bunching changed across different granularities of time. With the former, notably fewer bunching incidents occur on weekends relative to weekdays, with the lowest number of bunching incidents occurring each week on a Sunday (164 and 220 incidents for the first and second Sundays, respectively). Weekday incidents are uniformly high, with the second Friday (Jan 17th) experiencing the most amount of bus bunching across the study period. Nominally, the first and Wednesdays of the study period started may appear to be an exception in this trend, but is less likely an outlier and more likely the result of the fact that data collection started late and ended early on each of the days, meaning that the full seventeen hours of data was not recorded for either day (Fig 9). To capture the spatial distribution of where bunching occurs from day to day, a time-lapse heatmap of bunching incidents was also constructed. From the heatmaps we observe that across all day types, the Loop remains a place where bunching events tend to concentrate, albeit the absolute number of events is smaller during the weekends. Nevertheless, the ratio of bunching inside to outside the Loop remains mostly consistent. Other locations where bunching levels were consistently high include 79th street near Ashland Avenue, Chicago Avenue between Damen and Milwaukee, Clarke/North in Sedgewick, and Belmont/Clark in Lakeview. Since time lapse heatmap is not static, a link to the visualization can be founded in the Appendix section (Fig 7).

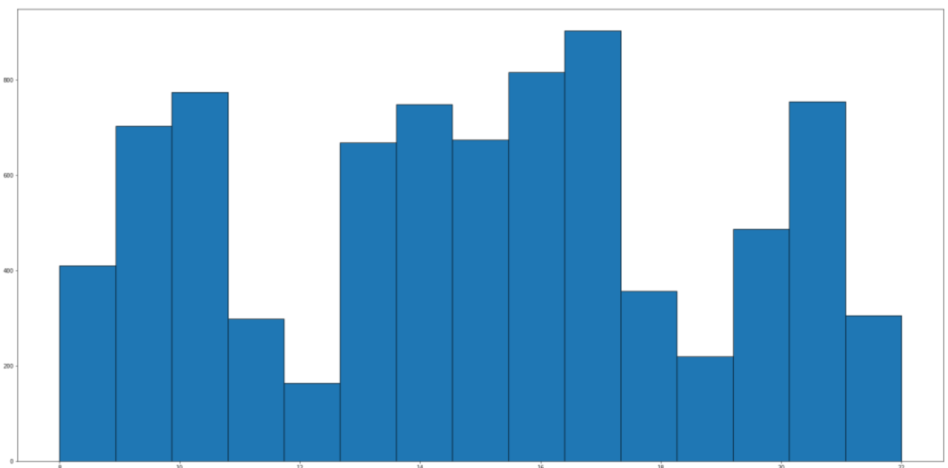

Fig 8. Bunching incidents, network-wide, by day

Aggregating incidents across by hour, a bimodal distribution of events is observed (Fig 9). Two peaks emerge with an earlier one at 8 am and another at 4 pm. These are consistent with morning and afternoon rush hour in Chicago, when traffic flows are at their highest. Interestingly, a much wider spread of events is noticeable around the after-work rush hour peak, while the first mode of the distribution at 8 am reaches its maximum with a much smaller spread relative to the times around it. Understandably traffic flows necessarily recede during the day while experiencing much larger marginal increases between early morning and the AM rush hour, which perhaps explains the difference in spread between the two peaks.

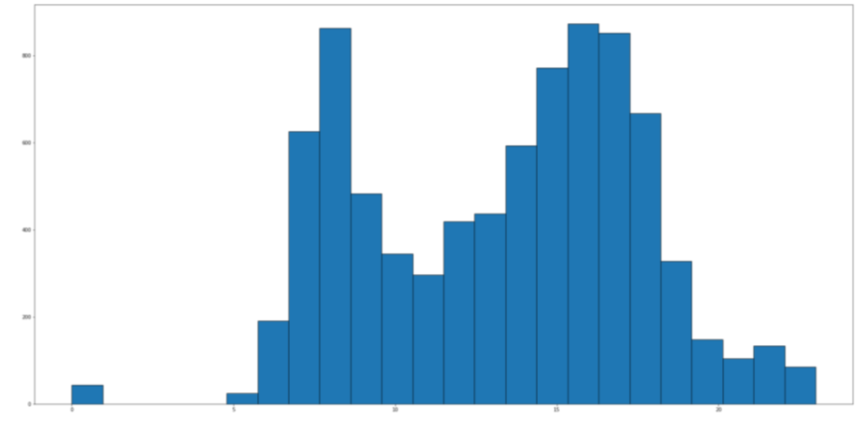

Fig 9. Bunching incidents hourly across network

Bunching at the Route Level. Increasing the granularity of observation, findings from analyzing bunching at the route level are presented next. To better understand the shape of the data, a first step involved ranking the routes by the number of incidents they experienced over the study period. This ranking accounts for both the number and duration of bunching incidents; this possible because the bunching identification method use tracks incidents at two-minute intervals, so longer duration incidents are reflected by the larger number of points that are relative to a shorter bunching incident. It is thus possible for a route to have a higher ranking than another if it encounters particularly long bunching incidents. All that said, the 79th St bus appears to have experienced the most bunching, followed closely by the 66 and 49 buses. On the opposite end, the 34, 55, 81 routes stand as the routes with the fewest number of bunching incidents over the period studied (Fig. 10).

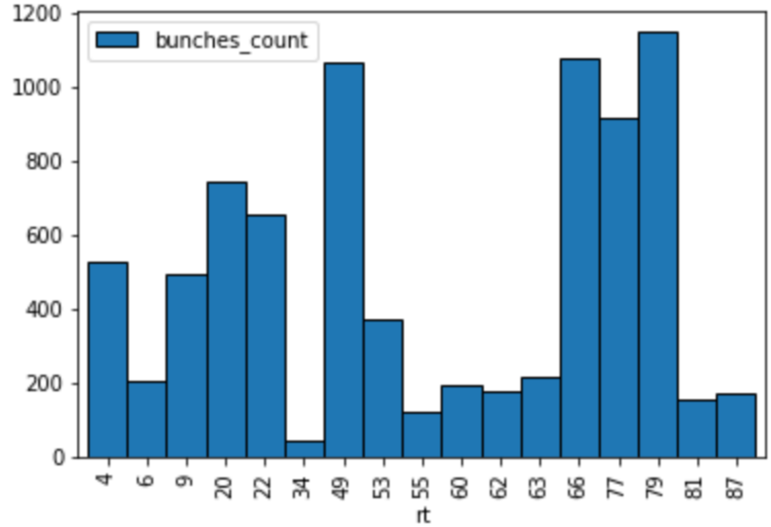

Fig 10. Bunching incidents by route

To contextualize these results, it is useful to recall the moderate, positive, relationship between service frequency, proxied by the average number of buses that run each day on a route, and delays. It is reasonable to expect some covariance between delays and bunching: a delayed bus is inherently an off-schedule one, meaning that headway deviance is inevitable, inviting a higher likelihood of bunching in the process (fitting the data indeed reveals a moderately positive relationship between delays and bunching, with a R-squared of .26). Here, with an R-squared value of .79, an even stronger positive correlation exists between service frequency with bunching: routes that have more buses running on them, experience more bunching incidents (Fig 11). In an attempt to normalize the number of incidents a bus route experiences agnostic of its service frequency, bunching incidents per route was divided by the total number of buses recorded on each route, giving the percentage of bus locations that had a bunching incident by route. Now, the rankings shift slightly. It emerges that route 20, while experiencing a small absolute number of bunching events, is ranked third in terms of share of bus locations received that turned out to be a bus bunching. The 79 bus drops to fifth place when bunching incidents are normalized by total number of buses running a route (Table 4)

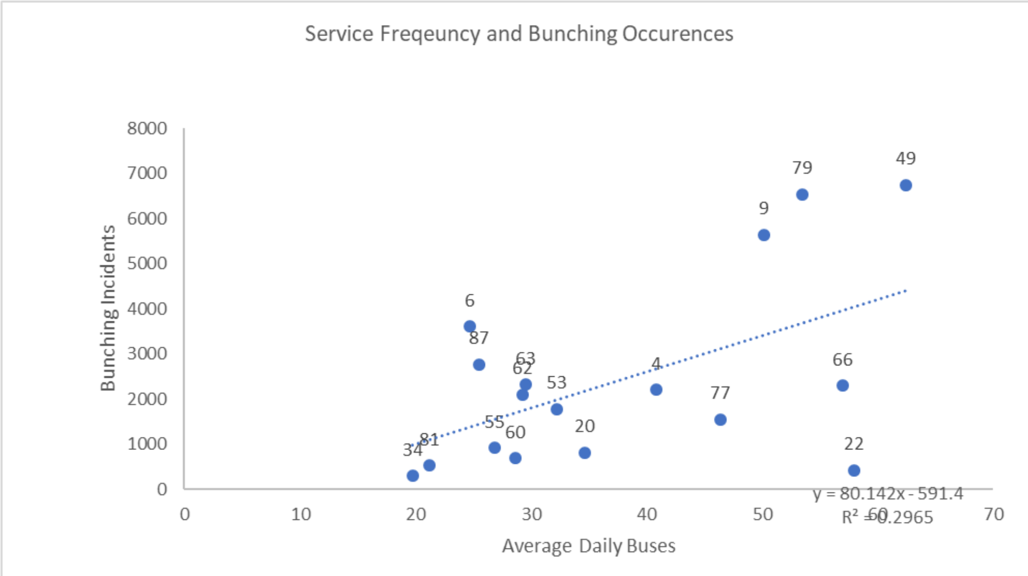

Fig 11. Relationship between Service Frequency and Bunching

rt	bus_count	bunches_count	ratio
49	138393	1062	0.007674
66	140850	1078	0.007654
20	100935	743	0.007361
77	133649	914	0.006839
79	170474	1148	0.006734
22	113955	657	0.005765
4	139003	525	0.003777
9	140625	494	0.003513
53	115979	370	0.00319
6	81790	207	0.002531
81	64797	153	0.002361
60	84312	197	0.002337
63	93547	218	0.00233
62	81998	179	0.002183
87	91977	174	0.001892
55	68367	121	0.00177
34	33181	45	0.001356

Table 4. Buses running to bunching ratio by route

In an effort to develop a more robust understanding of relative route performance that complements the single-variable metric calculated above, route bunching data is disaggregated further by time and route segment. The result histogram matrix reveals valuable new insights into bunching consistency along a given route (Fig 12). Most all the routes exhibit some degree of bimodality in distribution. Peaks in bunching incidents occur most frequently in the afternoon during the even rush hour. Some routes are highly bimodal around typical rush hour periods in the morning and evening and can thus be thought of as 'commuter' routes. These include the 9, 20, 22, 53, 55, 66, 77, and 79 buses. Other routes, while still bimodal, are wider in spread as bunching incidents remain relatively high even during non-rush hour periods throughout the day, suggesting that rush hour traffic explains less of the headway deviation that occurs on these more time-agnostic routes, which include the 4, 6, 49 routes.

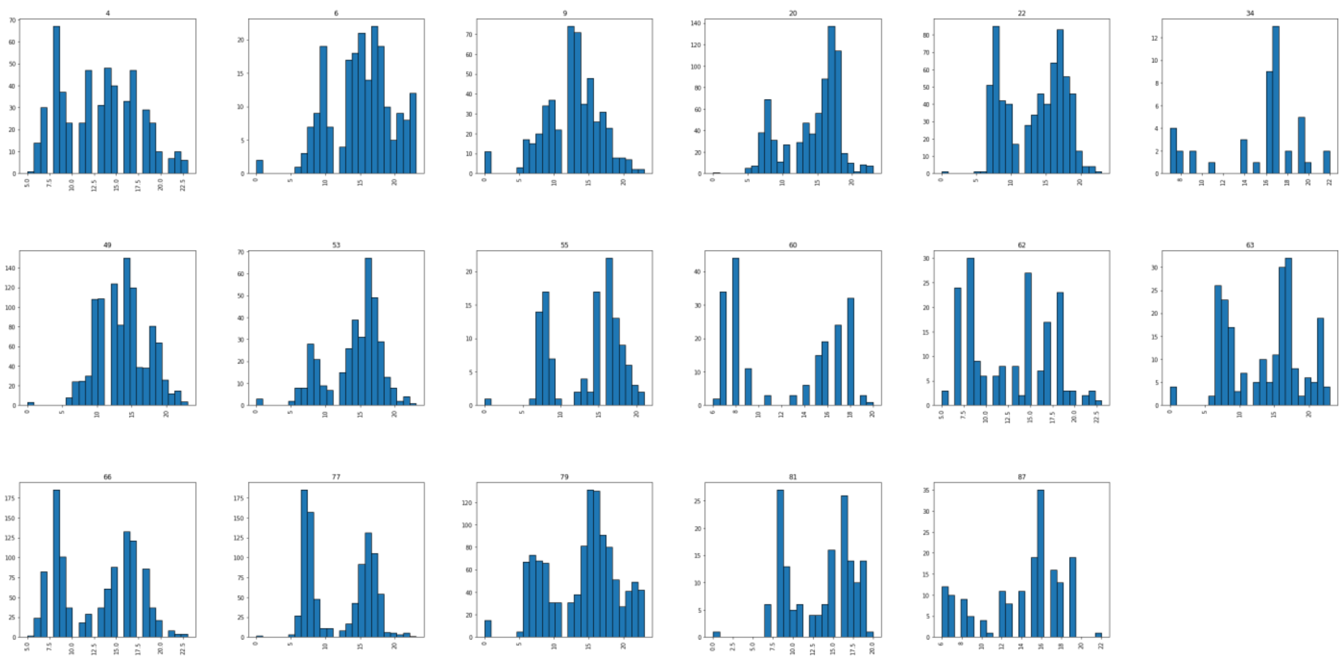

Fig 12. Distribution of bunching incidents over time, by route

Using the same temporal unit, routes were further differentiated by direction of travel, of which they are for, eastbound, westbound, northbound and southbound. Across all routes, buses traveling eastbound and northbound experience more bunching incidents than those traveling northbound and southbound. At the route level, these differences manifest as well. The 4, 49, and 79 routes exhibit particularly higher number of bunching incidents on their northbound and eastbound traveling buses relative to their buses traveling in the opposite direction. The 4, for instance, experiences more than six times the amount of bunching on its northbound route compared to its southbound route. The 79 is similar with nearly five times the number of bunching on its eastbound route compared with its westbound route (Fig 14). Separating bunching incidents by direction of travel has turned out to be an invaluable exercise, as transit planners can use the results to decide where to prioritize costly route improvements not only by route, but by direction of travel on the route as well. In this case, there is a discussion to be had concerning how street geometry on roads traveling north and east can be improved from transit service before similar discussions are had for roads leading in the opposite directions.

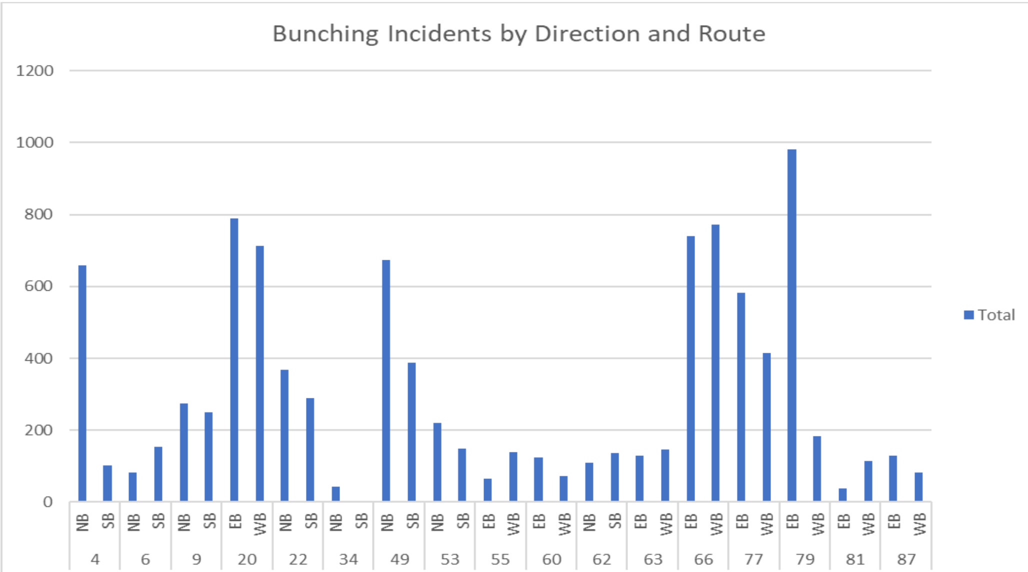

Fig 13. Hourly Bunching Incidents by Route

As for route segment analysis, owl bus routes were split into sections of equidistant length (500 meters). The route segments were subsequently buffered out into polygons so point-in-polygon analysis could be completed. Accounting for the presence of false positive bunching was particularly important for this part of the study, since an overestimated count changes the scale of bunching counts and increase likelihood of misidentifying route segments with high bunching incidence. As such, no bunching incidents within 200 meters of terminal points (derived from the average distance of a city block in Chicago being an eighth of a mile) were counted in the making of the route segment choropleth plot (Fig 14). The buffered route segments with counts of bunching incidents spatially joined serves informs the remainder of the discussion of spatial patterns of bus bunching in Chicago at the route level.

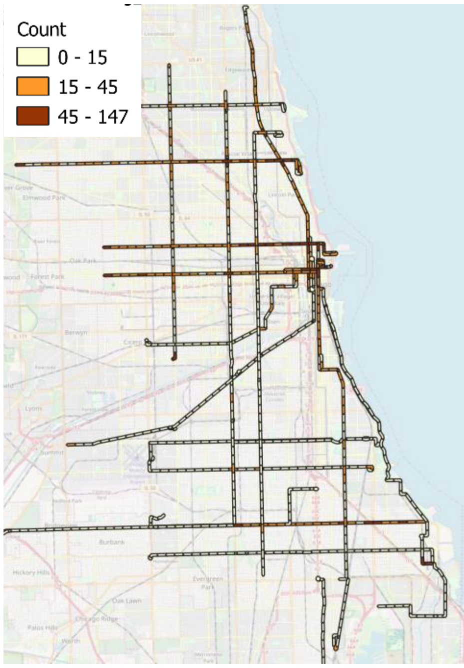

Fig 14. Bunching incidents per 500-meter route segments

Summary statistics of the route segments are displayed below (Table 5). The vast majority of segments experienced little bunching, but the higher mean compared to median value suggests that a few segments are severely affected by bus bunching, thus pulling the average value of incidents per segment higher than the median. From the standard deviation and the interquartile range, we can tell that segment incident values are relatively spread, again reflective of the fact that the most parts of CTA bus routes seldom face bunching, but the parts that do are particularly susceptible to it.

Statistic	**Value**
Number of Segments	880
Mean	13.8
Median	8
St dev	17.5
Q1/Q3	2, 18

Table 5. Bunching by Route Segment Summary Statistics

Turning attention now to the choropleth map itself, the general pattern that emerges follow what can be inferred from the data's descriptive statistics. Segments in the plot are classified using Jenk's Natural Breaks, which finds the optimal way to assign a range of values into classes such that variation is minimized between classes, for a given number of classes. In this application, I chose to divide incident counts into three intervals to identify route segments with 'high,' 'medium,' and 'low' numbers of bunching incidents. A plot of the entire mapping extent is provided above (Fig 14). From it, clusters of medium and high incidence segments are observable near the loop. Interestingly is the ability of this style of visualization to capture the change in the number of bunching incidents a route experiences as it approaches the Central Business District in the Loop. Filtering the plot for segments with the highest levels of bunching incidents results in a much sparser map. Three regions emerge as areas of interest in the North, Loop/Near North, and the Far South Sides (Fig 15). A thorough analysis of each area is conducted below.

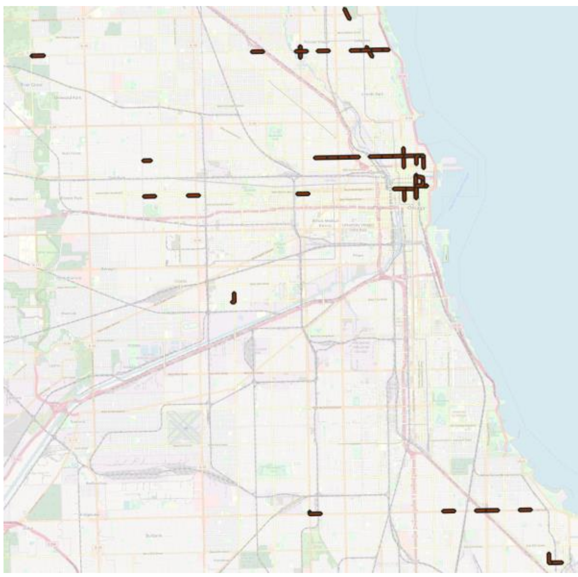

Fig 15. Route segments with high levels of bunching incidents

On the North Side, route segments with an elevated number of bunching incidents converge primarily in the North Center/Lakeview neighborhoods on Belmont Avenue between Kennedy Expressway to the west and Lake Shore Drive to the east (Fig 16). The longest continuous stretch of road here with high bunching extends for about 2,000 meters (~1 mile) between Lakewood Avenue and the lakefront on Belmont Avenue. In the greater scheme of things, this portion of Belmont Ave represents the last few stops on the eastbound 77 Belmont bus. Intersections between Belmont and other major mile roads in the area, Clark Street and Irving Park Road, are also highlighted in red as sections of the road with high levels of bunching.

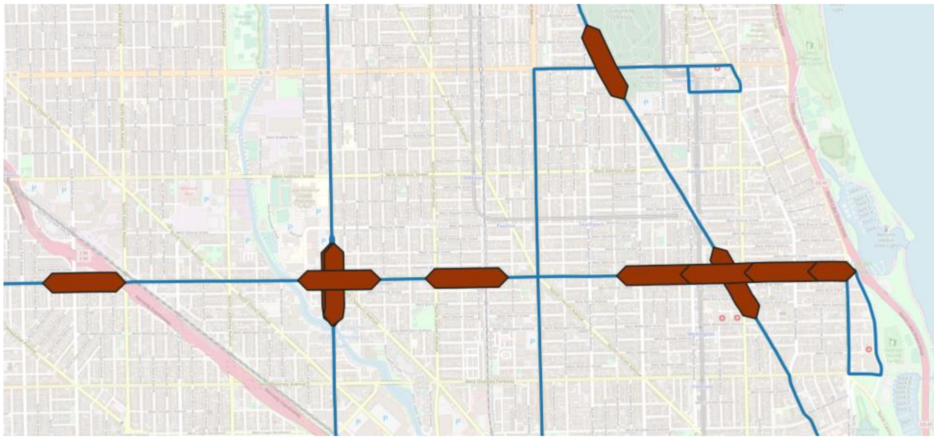

Fig 16. Route segment analysis, North Side

In the Loop/ Near North Side area, Chicago Avenue stands out with a conspicuously long portion of the 66 Chicago route being designated as a segment with high incidence of bus bunching (Fig 17). The section, which stretches from hallway between Western and Damen Avenue on the west side to past Michigan on the east side, represents a nearly 5,000 meters (3.1 miles) of continuously high levels of bunching. Chicago Ave's intersection with North Clarke Street is another pain point in the Near North Side as an area of high bunching concentration. Lower down in the Loop, bunching incidents continue on Clark Street between Washington and Jackson, further to the east, Michigan between Wacker and Adams was also identified as a location with high levels of bunching.

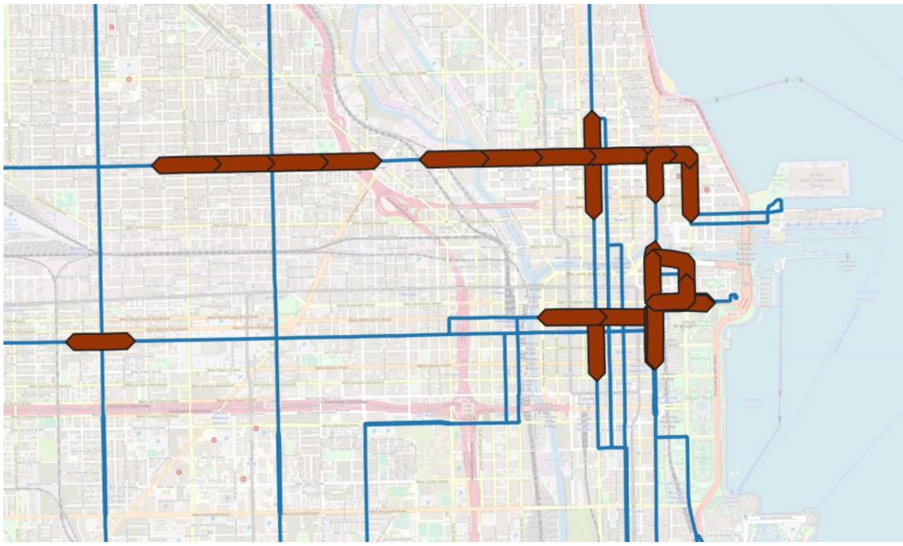

Fig 17. Route segment analysis, Near North Side/ Loop

Out of the three spaces identified to have some degree of concentration of route segments with high levels of bunching, the South Side are around 79th street displayed the smallest number of route segments designated as sections of road with high bus bunching (Fig 19). Here, the events most all take place on 79th street, which is traversed by the eaast/westbound orientated 79 bus. In addition two highlighted route segments at the intersections of 79th street with Ashland Ave (traversed by the 9 bus) and Cottage Grove (traversed by the 4 bus), sections of the 79 route that are also suceptible to bunching include the overpass transfer to the 79th St Red Line station and Yates Boluevard further east.

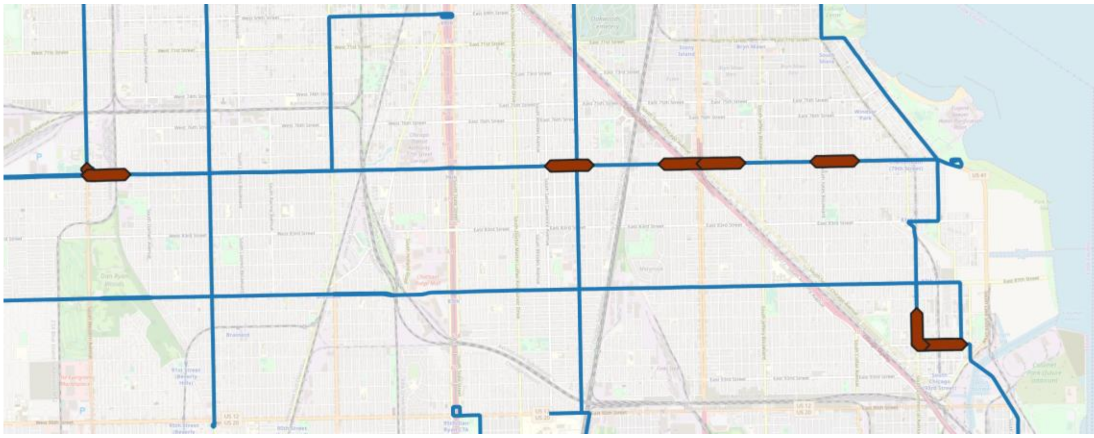

Fig 18, Route segment analysis, South Side

Bunching at the Stop Level. As previously discussed in the bunching identification section, the hope for analyzing bus bunching at the stop level is twofold. First, we seek to identify the bus stops—route agnostic—that tend to experience bunching more frequently relative to the average bus stop. Introducing stop locations into the calculation also avails the opportunity to explore new dimensions of the data, such as the relationship between the number of routes served at a stop and the number of bunching incidents it encounters. On applying the previously identified stop-level bunching identification method to the dataset of bus locations, care had to be taken to ensure that only buses and stops on the same route, in the same direction, were evaluated each iteration . As such the bus location dataset was filtered by direction of travel and the assignment function was applied four times to each cut of the bus data. Results were subsequently reaggregated to produce the figure below (Fig 19).

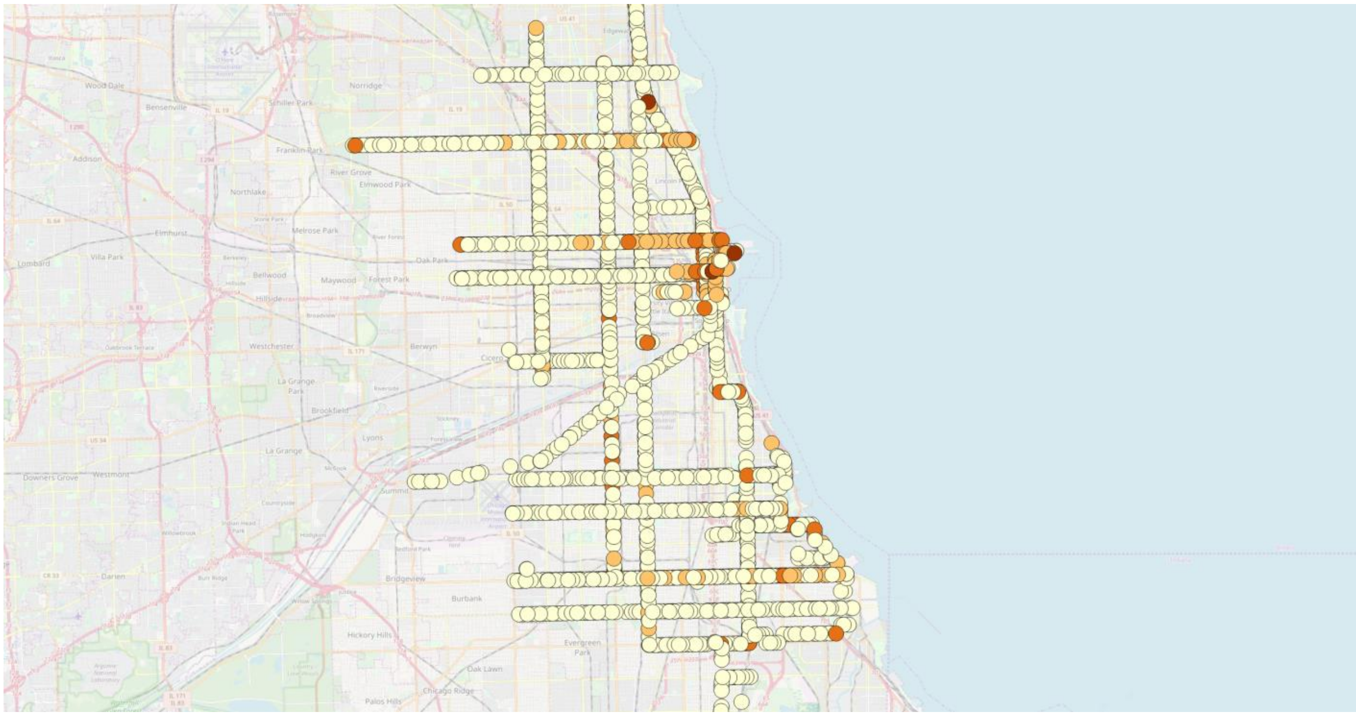

Fig 19. Distribution of bunching incidents by bus stop

The output produced is reminiscent of previous results from approaching the problem at the network and route level. Along most lengths of each route, bus stops received less than ten bunching assignments. Bus stops downtown and in a few other locations are an exception to this rule, creating a hugely left skewed distribution of bunching incidents assigned to stops. The Loop once again emerges as an area with concentrated bunching events with bus stops in the Loop receiving more than six times the number of bunching incidents compared with the network wide average (2.75 population average, 17.08 Loop average). Ranking all bus stops by associated bunching incidents reveals a collection of stops based in the Loop. One exception among them is the southbound stop at Western and 77th, which serves the 49 Western route.

PUBLIC_NAM	**ROUTESSTPG**	**DIR**	**Count**
Clark & Belle Plaine	9,22	NB	246
Western & 77th Street	49	SB	148
Illinois & Lake Shore	2,29,65,66,124	EB	124
Washington & Wabash	N4,J14,19,20,56,60,N66,124,147,151,157	EB	104
Michigan & Randolph	3,6,19,20,26,N66,124,143,151,147,157	NB	81

Table 6. Top 5 Bus Stops by Bunching Incidents Encountered

Since the initial bunching assignment function was run on the bus location dataset filtered by direction of travel, another useful comparison that could be quickly assembled looks at variances between bunching events experienced by bus stops stationed in opposite directions on each route, a granularity that is arguably more difficult to capture and depict when analyzing the data at the route level.

Fig 20, 21, 22, 23. NB, SB, EB, WB bus stops

Here, the difference in the directionality of bus bunching is evident, more so between westbound and eastbound buses than with northbound and southbound buses. At most, westbound buses were assigned nine bunching incidents, which would place them in the lowest Jenk class calculated for the set of eastbound bus stops. Particularly for eastbound and westbound buses in the Loop, the contrast between the bunching experienced between the two directions is interest and merits further analysis as to why, given that service frequency into the Loop should match quite closely with service frequency out of the Loop, such a stark difference exists. Comparing bunching incidents across bus stops in different directions also avails the opportunity to evaluate bus stops that may not encounter many bunching events relative to the population wide count but are of particular concern on a specific route. Sorting each set of bus stops by number of bunching events the stops were assigned generates a set of 25 bus stops that represent priority locations for future streetscape improvement projects carried out by the CTA.

PUBLIC_NAM	**ROUTESSTPG**	**Count**
Clark & Belle Plaine	9,22	246
Michigan & Randolph	3,6,19,20,26, N66,124, 143,151,147,157	81
Columbus & Randolph	4	61
Clark & Irving Park	9,22	42
Michigan & E. Wacker	2,3,,6,10,20, 26,N66,151,157	39

PUBLIC_NAM	**ROUTESSTPG**	**Count**
Western & 77th Street	49	148
Pulaski & 31st Street	53	32
Clark & Kinzie	22	19
Cottage Grove & 115th Street	4,111A,115	16
Western & 63rd Street	49,X49	14

Table 7,8. Northbound, southbound, top 5 stops with most bunching incidents

PUBLIC_NAM	**ROUTESSTPG**	**Count**
Madison & Wabash	J14,19,20,56, 60,N66,124,157	9
Madison & LaSalle	J14,19,20,56, 60,124,157	4
Madison & Franklin	J14,19,20, 56,60,124,157	4
Madison & Dearborn/State	J14,19,20, 56,60,124,157	4
95th Street & Langley	4,N5,95,115	4

PUBLIC_NAM	**ROUTESSTPG**	**Count**
Illinois & Lake Shore	2,29,65,66,124	124
Washington & Wabash	N4,J14,19,20,56, 60,N66,124, 147,151,157	104
93rd Street & Escanaba	N5,95	66
Belmont & Sheridan	77,151,156	66
Chicago & Fairbanks	3,26,66,157	51

Table 8, 9. Westbound, eastbound top 5 stops with most bunching incidents

Among the analyses that are unique to studying bus bunching at the bus stop level is the relationship between the number of routes a given bus stop serves and the number of bunching incidents it encounters. Intuitively, one might expect bus stops that encounter higher flows of bus traffic to be more susceptible to bus bunching. The collection of bus stops in the Loop would fall under such classification, for one. To explore this possibility more in depth, the length of the 'ROUTESSTPG' variable in the bus stops dataset was calculated in order to approximate a count of the routes that stop at each bus stop whose correlation with the count of bunching incidents was then estimated. Since no bus stop serves more than 12 routes (Michigan and South Water gets the honor of being that stop) and the majority of bus stops were assigned very few bunching incidents, the relationship between the two variables, while slightly positive, is dubious as best with a low R-squared value of .108.

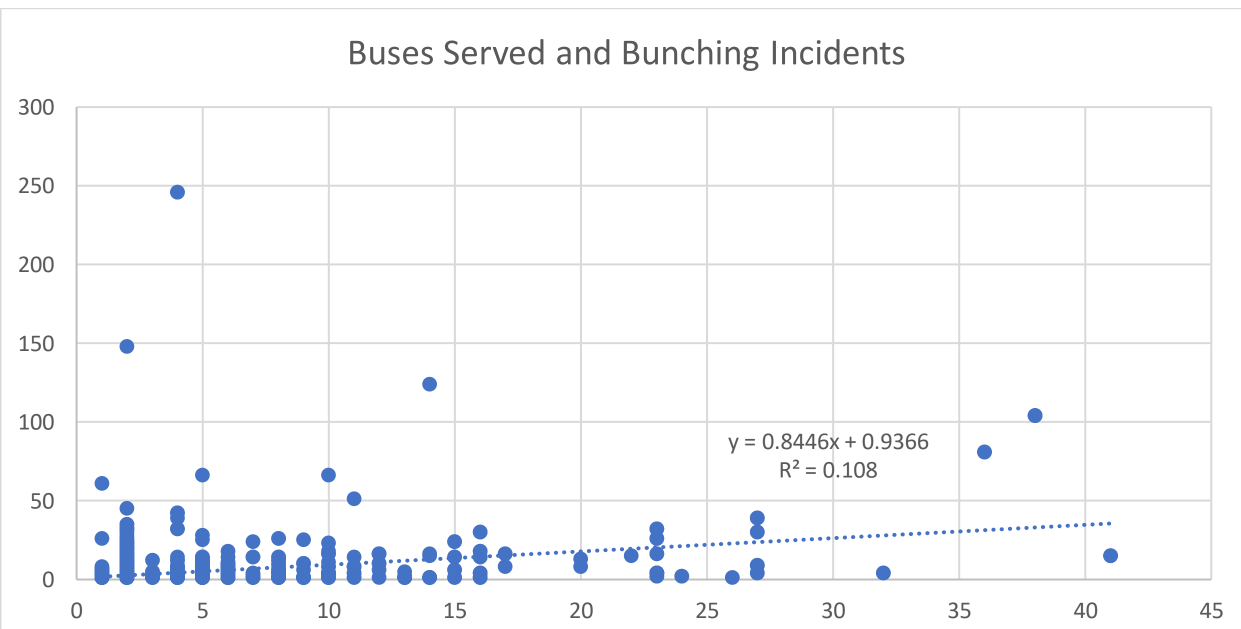

Fig 24. Correlation between number of buses served at a bus stop and bunching incidents

Prediction

Data Preparation. In constructing the predictive model for bunching at the stop level, data was collected and aggregated by bus stop. In this, count data of bunching incidents at the stop level had already been assembled for the analysis conducted in the previous section. Thus, much of the data collection, preparation, and processing discussed here relates to the work done around building out the data backing the model's potential predictor variables. As introduced in the earlier 'Methods and Data' section, a combination of features about the stop itself and the socioeconomic environment around it were proposed for the purposes of this study. With the latter, it is obviously the case that a stop has no 'population' or 'income' attribute. Instead, various socioeconomic indicators, procured from the data aggregator Social Explorer, were assigned to bus stops based on the level attained by the Census block group in which the stop is located. One variable, job density, was only available at the tract level but values of the variable were assigned to bus stops similarly. This was achieved using a series of table and spatial joins in QGIS. Several stops intersected block groups and/or Census tracts where no data was available. In these cases, an average of the socioeconomic indicator values from the surrounding tracts/block groups was calculated and assigned to the polygon with missing values.

As for the former category regarding features of the stop itself—where it is on the street, which direction of travel it serves, how many routes it serves, historical passenger boarding counts—most required little to no processing; direction of travel and position on the street could simply be labeled as a factor variable in R and were ready for estimation. Calculating the routes served by a bus stop was slightly more involved but more or less straightforward as the number of routes a stop serves, indicated by the 'ROUTESSTPG' column in the bus stop shapefile provided by the CTA, was summed for each row in the dataset in Python. Among the route attributes, historical passenger data at the stop level required the most attention during the data preparation period. The most involved part of this task involved reconciling historical passenger boarding data with more recent data on the location of relevant bus stops without consistent merge keys between the historical boarding data aggregated by stops and the larger bus stop dataset. For stops with historical boardings that could not be identified by merging the two data frames on their respective merge keys, nearest neighbors were calculated between the stops missing passenger counts and the historical passenger count dataset and the passenger count of the nearest stop (which often times *was* the stop of interest itself, simply with mismatching merge keys) was used.

The last part of data preparation involved constructing a lag term to control for autocorrelation and a rush hour indicator variable to capture the effect of high demand periods on bunching frequency. These variables were constructed by manipulating preexisting bunching and time data in Python. For the lag term, stops were sorted for a given route and direction in spatial order (e.g., northbound stops are ordered from the southernmost to the northernmost stop) and took note of bunching presentation at the previous stop. The lag variable takes a binary form, with true values indicating that bunching did occur at stop *s-1* for some given stop, *s.* The rush hour indicator variable is also binary, with bunching counts per stop being calculated for the morning/evening rush hour periods (set at the conventional 6-9 AM and 4-8 PM) as well as non-rush hour periods. An interaction term involving the rush hour and demand variable will be entertained in subsequent fittings of the model. With control variables added, the final dataset used in the training and validation of the predictive models consisted of 5,477 rows and 11 columns, where each row is a bus stop characterized by a unique set of route and socioeconomic features. A summary of the dataset's descriptive statistics, as well as a sample of the table rows, is included below for reference (Tables 10, 11).

STREET	CROSS_ST	PUBLIC_NAM	POINT_X	POINT_Y	DIR	POS	oct12_boardings	bg_percent_tr	bg_pop_den	ct_jobdense	lag_term	rts_served	bg_hhi	is_rush	bunching_counts
COTTAGE G	35TH STREET	35th Street & Co	-87.61108	41.83108	SB	NS	107.8	1	12300	298.329863	TRUE	1	30992.75	1	0
COTTAGE G	74TH STREET	Cottage Grove &	-87.60557	41.76053	SB	NS	38.2	0.434065934	13549.24	691.477962	FALSE	1	31050.33333	1	0
WESTERN	BERWYN	Western & Berw	-87.68934	41.97753	SB	FS	1047.5	0.278967254	38051.52	2437.50739	FALSE	2	67526	1	0
63RD STRE	CALIFORNIA	63rd Street & Ca	-87.69361	41.77919	WB	FS	227.3	0.25	23831.14	1692.05844	TRUE	1	23976	1	1
63RD STRE	FRANCISCO	63rd Street & Fra	-87.69569	41.77916	WB	NS	29.8	0.25	23831.14	1692.05844	TRUE	1	23976	1	0
63RD STRE	SACRAMENT(	63rd Street & Sa	-87.69819	41.77912	WB	NS	78.1	0.25	23831.14	1692.05844	FALSE	1	23976	1	1
63RD STRE	ALBANY	63rd Street & All	-87.70056	41.77907	WB	NS	38.9	0.25	23831.14	1692.05844	TRUE	1	23976	1	0
ASHLAND	65TH STREET	Ashland & 65th S	-87.66401	41.77576	NB	NS	47.7	0.210970464	11027.79	1056.30154	FALSE	1	32426	1	0
ASHLAND	GREEN LINE (	Ashland/63rd Gr	-87.66405	41.77887	NB	MB	597.5	0.210970464	11027.79	1056.30154	FALSE	2	32426	1	1
63RD STRE	ASHLAND	63rd Street & As	-87.66385	41.77941	EB	FS	889.3	0.210970464	11027.79	1056.30154	TRUE	1	32426	1	0
63RD STRE	LAFLIN	63rd Street & La	-87.66188	41.77945	EB	NS	70	0.210970464	11027.79	1056.30154	FALSE	1	32426	1	0
63RD STRE	LOOMIS	63rd Street & Lo	-87.65947	41.77948	EB	NS	101.9	0.210970464	11027.79	1056.30154	FALSE	1	32426	1	0
63RD STRE	GREEN LINE S	Ashland/63rd Gr	-87.66368	41.77881	WB	TERM	467.3	0.210970464	11027.79	1056.30154	FALSE	1	32426	1	2
COTTAGE G	75TH STREET	Cottage Grove &	-87.60549	41.75869	SB	NS	275.6	0.434065934	13549.24	691.477962	FALSE	1	31050.33333	1	1
DEARBORN	POLK	Dearborn & Polk	-87.62905	41.87251	NB	FS	230	0.366533865	101166.8	10755.7331	FALSE	2	114492	1	5
DEARBORN	POLK	Dearborn & Polk	-87.62905	41.87251	NB	FS	230	0.366533865	101166.8	10755.7331	FALSE	2	114492	1	5
S HYDE PAI	54TH STREET	S Hyde Park & 54	-87.58395	41.79835	NB	FS	192.7	0.387878788	32200.01	2706.08925	FALSE	3	55286	1	0

Table 10. Table for regression analysis

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
oct12_boardings	5,476	173.644	296.376	0.000	30.600	187.300	3,366.300
bg_percent_transit	5,476	0.289	0.148	0	0.2	0.4	1
bg_pop_den	5,476	16,678.020	15,054.180	349	7,493.4	22,001.9	238,747
ct_jobdense	5,476	22,505.590	97,303.400	11.067	912.969	5,841.248	729,852.300
rts_served	5,476	1.688	1.440	1	1	2	12
bg_hhi	5,476	57,959.580	35,489.800	5,846.000	32,266.000	78,047.000	214,659.000
bunching_counts	5,476	1.500	9.436	0	0	1	348

Table 11. Summary statistics

A correlation matrix of the numerical predictor variables is also presented (Table 12.). The matrix suggests some positive relationships do exist, particularly among the socioeconomic ones. Since the coefficients are relatively small though, it is still unclear though at this point whether the degree of correlation is enough to engender multicollinearity concerns in these predictor variables for the models built later on. After conducting these initial explorations of the data, the final step in this stage involved splitting the data into train/test sets for use in subsequent model fitting and prediction. A conventional 80-20% split was used.

	oct12_boardings	bg_percent_transit	bg_pop_den	ct_jobdense	bg_hhi	reg.rts_served
oct12_boardings	1	0.081	0.093	0.343	0.135	0.490
bg_percent_transit	0.081	1	0.182	-0.034	-0.039	-0.004
bg_pop_den	0.093	0.182	1	0.091	0.233	0.094
ct_jobdense	0.343	-0.034	0.091	1	0.249	0.499
bg_hhi	0.135	-0.039	0.233	0.249	1	0.284
rts_served	0.490	-0.004	0.094	0.499	0.284	1

Table 12. Predictor variable correlation matrix

Model Selection. Four models were fitted during model selection: Ordinary Least Squares (OLS), Poisson, NB, and hurdle NB (see appendix, Table 13). For selecting models, every proposed predictor variable was included in the model, with the intent being to prune variables after a model type was chosen. The reasons for OLS' inappropriateness here are manifold, but a model of its type was fitted nonetheless to set a baseline. Diagnostics indeed confirm the need for a model that can accommodate count data. With an R2 of just .01, the OLS model explains next to nothing about the variance in the bunching count data.

The remaining three models fitted are all more geared towards count data, which constitutes the response variable being predicted in this study. Count data consists of discrete, non-negative integer values (0,1,2,3….) and tends to contain many zeros in distribution. As such, they readily violate, among other restrictions, the normality assumption of OLS models and thus require different handling during estimation. Poisson regressions are commonly used to model count data and was thus the first choice for fitting here[51]. Poisson regressions assume a Poisson distribution of data with the form $\Pr(Y = y|\, \lambda = \frac{e^{-\lambda}\, \lambda^{y}}{y!}$, where lambda is both the mean and variance of the Poisson distribution. Instead of OLS estimations, count regressions calculate model coefficient using maximum likelihood estimation (MLE). While Poisson models do indeed appear to better reflect the distribution of count data, the (conditional) mean-variance equality implied in its probability distribution is difficult to satisfy when working with real world data. More often than not, real world count data tends to be over-dispersed, that is, the variance is greater than the mean. To investigate this possibility, I began by generating some descriptive statistics, specifically, the within level (in this case, the 'DIR' variable), and thus conditional, means and variances of bunching counts. The inequality between the calculated values certainly points towards over-dispersion in the response variable. After proceeding with the Poisson regression, a more formal hypothesis test conducted using the dispersiontest() function in the AER package in R provides more compelling evidence of over-dispersion (Figure 25.). In the event of over-dispersed data, negative binomial regressions become relevant as the assumptions of mean-variance equality are relaxed in negative binomial distribution. When binomial distribution is assumed, $E(Y) = \mu$ and $Var(Y) = \mu + \frac{1}{r}\mu^2$, where r is some real-valued parameter. When r is sufficiently small, negative binomial mean and variance are not equal, which is useful when modeling over-dispersed count data, as is the case here.

	M	**SD**
EB	2.02	11.69
NB	1.86	11.54
SB	1.19	6.29
WB	0.92	2.57

```
        Overdispersion test

data:  p_model
z = 2.1245, p-value = 0.01681
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  46.94827
```

[51] Chatterjee, *Regression Analysis by Example*, 360.

Table 14, Fig 25. Within level mean and standard deviation, Overdispersion test

Another feature to account for in the bunching count data is the large number of zeros in the dataset. Comfortingly, bus bunching did not appear to widely effect every bus stop during the period studied. In fact, the majority of bus stops did not experience bus bunching at all. For modelling purposes though, this presents a problem since there is more zeros than expected under a Poisson or even a negative binomial distribution. Typically to deal with 'excess' (in the sense that there is more than should be given a chosen distribution) zeros, a zero-inflated Poisson or NB model is used to fit the data[52]. Zero-inflated models assume that there are two zero generating processes explaining the number of zeros observed[53]. One generates zeroes as a part of the count process, the other represent absolute zeros. To concretize this concept, imagine collecting data on the number of shots made by basketball players during a game. Some players attempt a shot but are terrible at shooting and make zero baskets. Other players do not shoot at all. Here, the zeros emerging from the bad shots of the team follow a count process, while the zeros from the players who made no attempt to shoot at all are the product of another zero-generating process. With zero-inflated models, the assumption is that a part of the 'excess' zeros belongs to this second process. In the case of bus bunching, it was difficult to devise an underlying theory about bus stops that would suggest the way they experience bunching incidents is subject to two zero generating processes; bus stops possess no attributes that would make them *always* experience zero bunching incidents. As such, even though zero-inflated models were proposed in the methods section of this study, a hurdle model was fitted instead given our inability to identify two zero generating processes for the number of bunching incidents experienced by a bus stop.

Hurdle models are similar to zero-inflated models in that they also split the fitting of count data into two equations. Where hurdle models differ is in how the data is split. With hurdle models, there is no assumption of two zero generating processes. Rather, a probability model, commonly a logit, is calculated for the zeros in the

[52] Rodríguez, German. "Models for Count Data with Overdispersion," (Nov 2013). https://data.princeton.edu/wws509/notes/c4a.pdf
[53] *Ibid,* 5.

data while the remaining non-zero values in the data are fitted using a 'truncated' count model[54]. In other words, the two value generating processes in a hurdle model are one that produces zeros and one that produces positive values, which better matches the context of this study. Furthermore, by design, the hurdle model will perfectly fit zero counts in the training dataset, which helps ensure that zeros are not underestimated even after assuming NB distribution of the data. Given prior identification of over-dispersion in the data, a truncated NB model was selected to fit the stops with positive bunching counts. It should be noted that when fitting the hurdle model with all the proposed predictor variables, NA's in the diagnostic values of the model summary were returned, suggesting collinearity in the predictor variables, or (somewhat less applicable), too many predictor variables for the sample size. The hurdle model presented in the model comparison table thus reflects the removal of the predictor variables with NA's in the initial fit.

Comparing model diagnostics across the Poisson, NB, and hurdle NB models involved plotting rootograms of the three models, as well as checking the more conventional measures of relative fit, the Akaike Information Criterion (AIC). With AIC's of 26751.50, 12615.99, 12090.80 respectively, it is clear that fitting the data using a NB regression as opposed to a Poisson regression resulted in substantially improved fit, while just marginal improvements in fit were generated from adopting a Hurdle NB model. Plotting the rootograms of each respective model corroborates this observation (Figure 26). The Poisson model expectedly demonstrates underestimated zero incident counts and overestimated nonzero counts. The NB and hurdle NB models are visually similar, with the only overtly discernable difference being the relatively better fit of small incident values in the hurdle model. Given these diagnostics, the NB and hurdle models were chosen to predict bunching counts at the stop level in the test dataset. First, they were subject to a series of variable pruning procedures.

[54] *Ibid*, 6.

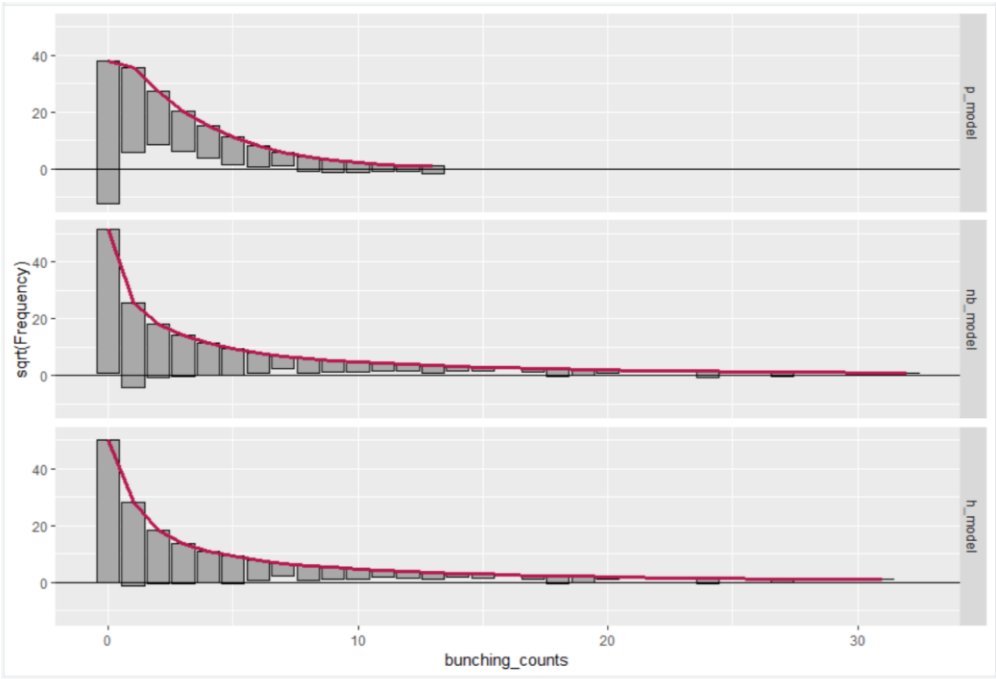

Figure 26. Count regression model diagnostics

Variable selection and prediction. While researching best approaches for variable selection, I was greeted with much caution regarding the implementation of stepwise variable selection procedures. Common reservations about BE and automatic variable selection procedures like it include the exclusion of theoretically important, but statistically insignificant variables, or the inclusion of circumstantially significant variables.[55] It should be noted that these concerns primarily relate with the ability of a model to conduct inference. The worry is that a mechanistic additional/removal of variables obscures or omits variables that motivate the underlying theory of the model. The primary objective of this analysis, however, is prediction, not inference. It feels relatively more reassuring to entertain some stepwise procedures as a result. Also, more well-received methods of selecting variables (e.g., ridge regression, LASSO) have yet to be implemented in R for the types of models, particularly the hurdle model, considered here. I thus defer back to the more established form of variable selection for the hurdle model.

[55] Smith, Gary. "Step Away from Stepwise." *Journal of Big Data* 5, no. 1 (2018). https://doi.org/10.1186/s40537-018-0143-6.

Within the set of mechanistic variable selection procedures, BE was chosen for more flexibility in inspecting the full equation as well as for being able to handle collinearity better than forward selection[56]. On the issue of collinearity, VIFs calculated for the initially fitted models, both NB and hurdle, were close to one, indicating low multicollinearity:

	GVIF	Df	GVIFDf))
oct12_boardings	1.102	1	1.050
DIR	1.381	3	1.055
POS	1.393	6	1.028
lag_term	1.158	1	1.076

Table 15. Hurdle model VIFs

	GVIF	Df	GVIFDf))
oct12_boardings	2.552	1	1.598
is_rush	2.502	1	1.582
bg_pop_den	2.250	1	1.500
bg_percent_transit	1.083	1	1.040
ct_jobdense	1.408	1	1.187
bg_hhi	1.197	1	1.094
DIR	1.076	3	1.012
POS	1.137	6	1.011
rts_served	1.838	1	1.356
lag_term	1.039	1	1.019
oct12_boardings:is_rush	2.445	1	1.564
is_rush:bg_pop_den	3.402	1	1.844

Table 16. NB model VIFs

Now, the primary concern addressed by iterating through the stepwise selection process is possible overfitting. In the hurdle model, this is less of a worry since the variables have already been reduced from the full set of possible predictors. The backwards elimination algorithm implemented through the step() function in R confirms this as no variables were recommended for removal in the hurdle model. The NB model went through no such previous pruning and was thus subjected to two types of variable selection: BE and LASSO. All in all, four models were

[56] Chatterjee, *Regression Analysis by Example,* 309.

selected for predicting on the test data: the previously reduced hurdle model, the NB model with full predictors, the BE NB model, and the LASSO NB model. Applying BE procedures on the NB model removed the ‘ct_jobdense,’ ‘bg_percent_transit,’ ‘is_rush,’ and ‘oct12_boardings’ interacted with ‘is_rush’ variables from it. Using the glmregNB() function in mpath, LASSO coefficients were estimated for the NB model as well. Comparing AIC’s across the three variations of the NB model, AIC of the BE NB model slightly improved relative to the full NB model (12611 vs 12616). Smallest AIC reported from LASSO model was found to be the same as the full NB model, 12616. Overall, the hurdle model still appears to perform better than the NB model, as summarized by the table below. To compare performance of the LASSO NB model to its peers, testing on out of sample data was conducted.

	Model	**Type**	**AIC**	**BIC**	**RMSE**	**Performance_Score**
1	h_model	hurdle	12,090.800	12,250.420	2.630	0.600
2	be_nb	negbin	12,611.260	12,726.190	0.903	0.411
3	nb_model	negbin	12,615.990	12,750.070	0.903	0.400

Table 17. Performance comparison

The predict() function in R was used to generated predicted bunching counts for a given stop in the test dataset (‘stops_test’). Model performances, that is, how accurately bunching counts per stop was predicted, using three approaches. The first of these was calculating the root mean squared error (RMSE) between actual counts and predicted counts for a given bus stop in the test dataset. RMSEs for the four models were all around ~11 (Table 18), suggesting that the modifications that improved model fit on the training data did not necessarily transfer to improvements in fit on the test data. Interpreting the RMSE is a bit tricky, since the values are based on discrete count data as opposed to continuous data as is usually the case when calculating RMSE. Here, it is likely that the value of the RMSE is inflated since it is sensitive to outliers, which do exist in the test data. The magnitude of the RMSE follows the unit of the dependent variable, which in this case is bunching counts.

Hurdle	NB	LASSO NB	BE NB
11.3207225530921	11.3373132413657	11.331839389155	11.3361864878623

Table 18. RMSEs, errors in predicted vs actual bunching counts

In addition to calculating RMSEs for each model, predictive accuracy was visualized using two methods. One simply involved comparing the total number of bunching counts predicted by each model to the actual total incidents across the entire testing dataset (Figure 27). Aggregating predicted values in this manner highlighted the underprediction of bunching by the LASSO NB model. This is interesting to consider given the similarity in RMSEs between the four models. A more involved visualization of the predictive accuracy of the model involved identifying 'hot' and 'cold' clusters of bunching incidents using the Getis Ord statistic (Figure 28). Comparing the hot and cold spots identified using observed data with the hot and cold spots identified using estimated data highlights spatial components of the prediction previously obfuscated in non-spatial treatments of the results. For one, the cluster maps show how efforts thus far to account for the large number of zero counts in the dataset may be insufficient—many more significant low value clusters are identified when overlaying observed data. On the other hand, clusters with high counts were more readily identified, again corroborating this notion that the models fitted were unable to adequately account for the heavily right skewed distribution of small values in both the training and the test data.

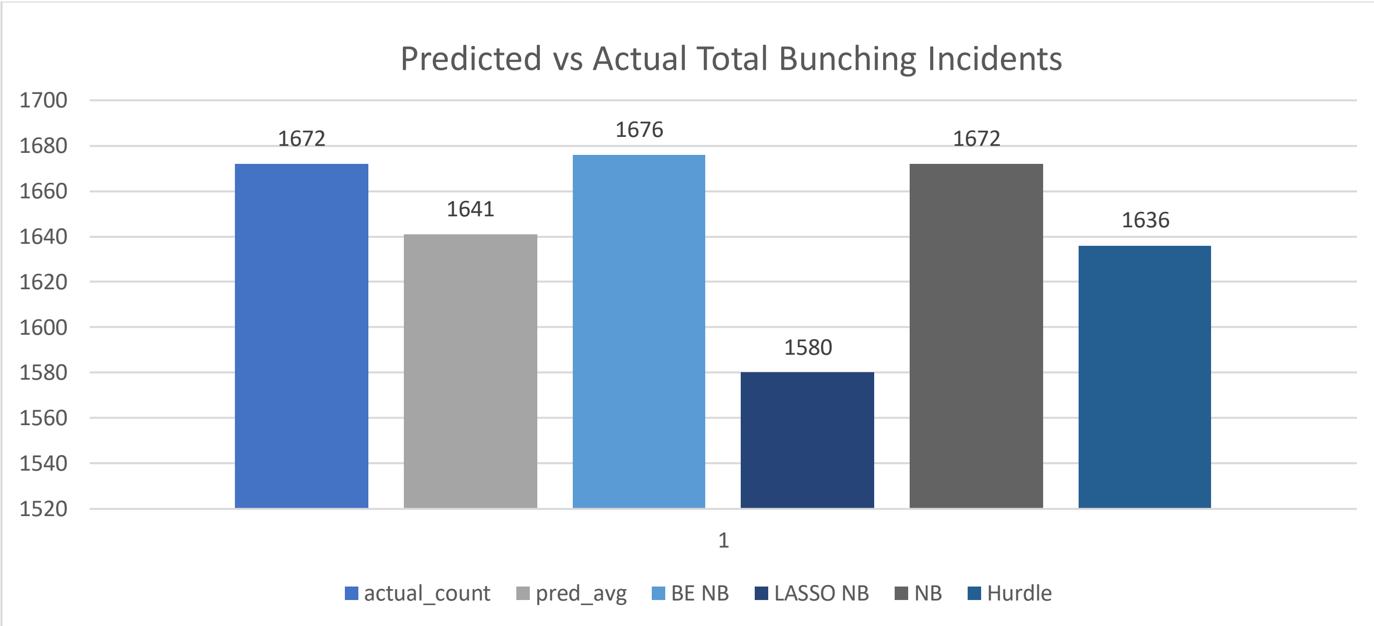

Figure 27. Predicted vs Observing bunching incidents by model

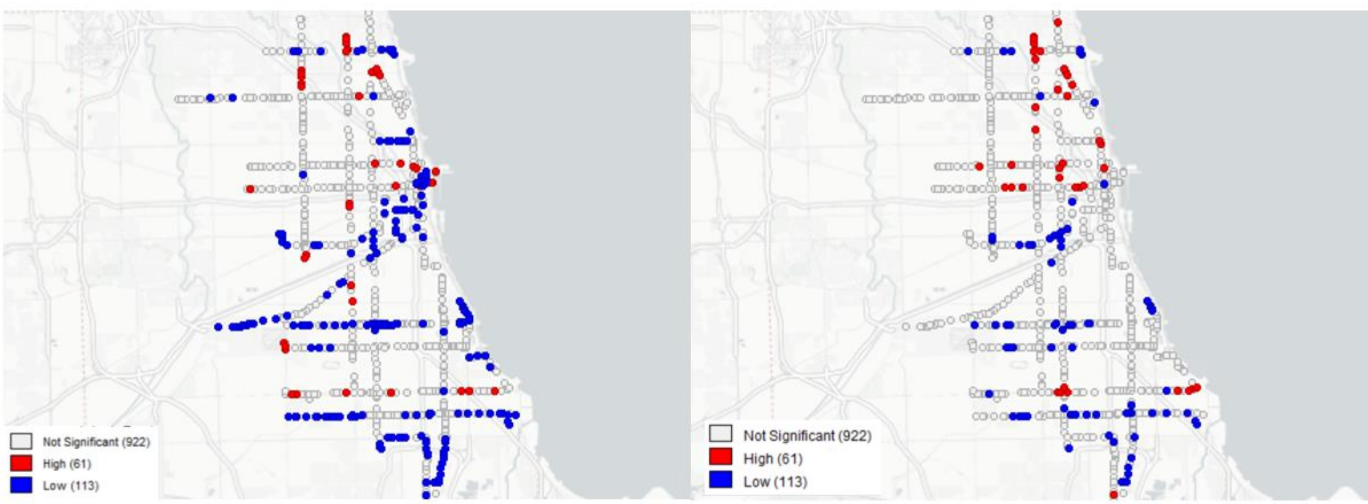

Figure 28. Bunching cold and hot spots, observed (left) vs predicted (right)

Conclusion

Discussion of results. To better understand the particular form of bus bunching in Chicago, research conducted for this paper sought to study the spatiotemporal patterns of bus flows in the CTA network. The results of data analysis for the most part reinforce prior research and personal intuitions about where and when buses tend to bunch along CTA routes. Using data gathered from a core network of CTA bus routes, the analysis of the spatial and temporal distributions of bus bunching in this study indicates that bunching occurs more frequently during morning and evening rush hour periods and in areas where inter-route interaction is high. A higher prevalence of bunching for routes when heading eastbound and northbound rather than southbound and westbound was observed as well.

This study also entertained the question of where bus bunching is likely to occur in the future. As stated previously, the primary goal of regression in this study is prediction. That said, it would be an opportunity remised to not extend the most cursory of inspections to the variables used in the various models estimated. In doing so, it is important to remember that underlying all these efforts is the fact that all the models constructed were, in absolute terms, weak fits of the data. As discussion of these models takes a more inferential tone, the conclusions drawn should be taken with more than a grain of salt. Moreover, in some cases, take the LASSO NB model or the BE NB model for example, the transformations applied to the model in an effort to optimize predictive performance preclude a coherent interpretation of the coefficients.

More can be said about the pure NB and hurdle models. Looking first at the pure NB model, we note that many of the highly significant coefficients fall into the 'route feature' category of variables. Indeed, this is a trend that is repeated across most all models fitted. Interpretation of these route feature coefficients vary. Within the 'DIR' and 'POS' variables are four direction levels in the former and seven stop position levels in the latter. As a log-linear model, coefficients must be exponentiated to obtain their direct effects on Y. The NB model for example, suggests that the baseline expected bunching count at a bus stop is about 0.7. Being a terminal stop increases average bunching incidents there by about 3 times. In the same model, median household income is one of the only socioeconomic variables that was found to be significant. The SE variables largely have minimal multiplicative effect on bus bunching at stops. This result picks up on an important pattern about socioeconomic variables and their effect on bus bunching, which is to say that they contribute little to explaining bunching occurrences at all—in the hurdle model, they do not even appear as parameters. Again, remaining cognizant of the weak absolute fit across the models tested, this result is helpful in informing the next iteration of model construction; socioeconomic variables need not be so heavily emphasized, route feature perhaps more so—particularly if the goal this time around is to model the data generating process rather than predicting future values.

Some more in depth reflection on the crux of this paper's analysis—prediction locations of future bus bunching—is entertained next. Before even considering test data and predictions, it was clear that there were going to be some issues with accounting for trends in the bunching data. Some of this extends from the bunching data itself. Its over-dispersion (more than half the stops had no bunching incidents) and long right tail (but a few had hundreds) present a challenging distribution for most any modelling approach to capture. The clusters of high low bunching stops depicted by the Getis Ord maps perhaps capture this fact better than any other manipulation of the data. The simultaneous under-identification of bunching 'cold spots' and relatively higher fidelity identification of hot spots suggest that the models, even those specifically designed to estimate the zero counts in a distribution of response variables, were lacking in some predictive power.

The other part of the challenge of prediction for this study came from the explanatory variables. Certain attributes found to be instrumental in determining the form of bus bunching in a given environment were unavailable; access to only publicly available data decreased this study's ability to obtain a more robust understanding the data generating process behind bus bunching. This effect was amplified by a desire to produce outcomes that would be easy to receive, digest, and act on from a policymaking standpoint. While designing this analysis, much concern was afforded to the applicability of public policy—the thought was to predict bunching incidents wherever it happens is not so useful for transportation authorities like the CTA, who might want to address the most palpable failures in their system. Concertizing this, modelling bunching that occurred during the express part of a route where buses are not stopping seemed much less valuable than modelling bunching at the bus stops, where customers actually interact with the bus network most directly. This created limitations in the data that could be used to model the bunching phenomenon that was again, further amplified by an inability to obtain certain types of data that previous literature has demonstrated to be pivotal in estimating the behavior of bus bunching. Overarching all of these considerations is an increasingly believable possibility that bunching does not follow a linear process at all. Bunching as a non-linear process would violate perhaps the fundamental assumption of linear regressions, that is, that variables enter the model linearly. Future efforts to predict where bus bunching should carefully consider this possibility, perhaps placing greater importance on using simulations and nonparametric methods to accurately identify determinants and predict trends. Below, I present my policy recommendations and offer thoughts on where to take research on this subject next.

Policy recommendations. This study makes three recommendations based on well-studied transit planning strategies that reliably improve bus on-time performance. These strategies are bus rapid transit (BRT), dedicated bus lanes, and transit signal priority (TSP). In this section, I first give a broader overview of what these options entail and what evidence exists to substantiate their effectiveness. I then contextualize these recommendations using the findings of this study. Specifically, three policy recommendations are offered: (1) Reviving BRT along the Ashland Avenue corridor, (2) creating new, or better enforcing, dedicated bus lanes on the 49, 66, and 79 routes, and (3), expanding TSP installments at major street (arterial streets typically located at mile/half-mile intervals) intersections.

Bus Rapid Transit. The Institute for Transportation and Development Policy describes BRT as a "high-quality bus-based transit system that delivers fast, comfortable, and cost-effective services at metro-level capacities. It does this through the provision of dedicated lanes, with busways and iconic stations typically aligned to the center of the road, off-board fare collection, and fast and frequent operations."[57] Essentially, BRT bus routes function like on-road rail systems. Among its key features is that buses running a BRT route are isolated from private vehicle traffic. Instead of being a part of the larger traffic flow, buses get their own lane, typically on the median of the

[57] "What Is BRT?" Institute for Transportation and Development Policy. ITDP. Accessed February 9, 2020. https://www.itdp.org/library/standards-and-guides/the-bus-rapid-transit-standard/what-is-brt/.

road, rather akin to how CTA Red and Blue Line trains are positioned between the northbound and southbound lanes on the Dan Ryan Expressway, albeit at a much smaller scale between arterial roads as opposed to a highway.

Fig 27. Rendering of a proposed BRT corridor on Ashland Ave in Chicago, IL

Another similarity between BRT and light/heavy rail transit in cities is that riders prepay to ride. At each station, fare booths are installed, and riders purchase tickets prior to boarding, which is typically done at the time of boarding, on the bus, on non-BRT routes. BRT is designed to lower the number of variables buses are subject to while running a route, thereby increasing their increasing probability of schedule adherence. While reducing the number of deviations from scheduled headways is arguably the most relevant result for a study on bus bunching, BRT has also been found to increase the operational speed of buses, ridership numbers, and decrease transportation related emissions. [58] In transit systems that have implemented it, riders saw decreased travel times of as high as 55%, and the city experienced many ancillary benefits (increased property values in land near BRT, tax revenue, etc) as well. [59] Since transit authorities can rely on existing capital inventories of road infrastructure and bus

[58] Goodman, Joesph, Melissa Laube, and Judith Shwenk. "ISSUES IN BUS RAPID TRANSIT." Federal Transit Administration . Accessed February 9, 2020. https://www.transit.dot.gov/sites/fta.dot.gov/files/issues.pdf.
[59] 'Bus Rapid Transit: Elements, Performance, Benefits." US Department of Transportation Federal Transit Administration. https://www.transit.dot.gov/sites/fta.dot.gov/files/BRTBrochure.pdf

inventories to support BRT, funding, planning, and constructing new BRT lines is much less costly relative to building out new rail-based transit lines[60].

Discussions of developing BRT corridors in the CTA bus network have been ongoing since at least the early 2010s. In 2013, the CTA publicized plans for a BRT line to be installed on Ashland Ave between Irving Park and 95th St, the terminuses of the 9 Ashland bus. The CTA cited reasons manifold for a BRT on Ashland, from high ridership on the corridor to the existing road infrastructure's ability to accommodate road median development, to Ashland Ave's position as a transfer corridor to many other CTA routes, rail and bus.[61] On performance improvements, the authority estimated that operational speed for buses running on Ashland could increase as much as 83%.[62] After completing an environment assessment for the project though, further plans have stalled as the CTA faced strong push back from community stakeholders who feared that the elimination of street parking on Ashland and the changes in traffic patterns would negatively affect businesses along the corridor. Although the 9 route experienced only moderate amounts of bus bunching relative to other routes in this study, a BRT line on Ashland is well-positioned to not only decrease schedule deviation on the 9 Ashland bus, but also reduce interaction effects the route has with some 37 other bus routes that intersect the line.

Dedicated Bus Lanes. Another option for improving service reliability on the CTA bus network is to implement a key feature of f BRT systems, dedicated bus lanes (DBL). DBLs represent an economical and efficient second-best alternative for minimizing the amount of on-route variance buses encounter while running a route. While all BRT systems have dedicated bus lanes, not all dedicated bus lanes, that is, sections of the road partitioned exclusively for bus transit, are BRT. BRT encompasses not only DBLs but other transit infrastructure as well, like fare stations for prepaid boarding and designated boarding stations. Moreover, while BRT lines typically have well-defined barriers from other traffic, DBLs are often demarcated less formally, usually with painted lanes or traffic cones. As a result, DBLs are susceptible to unauthorized use—private vehicles using the lane to bypass traffic, or

---

60 *Ibid.*
61 "Why Build BRT on Ashland?" Chicago Transit Authority.
https://www.transitchicago.com/assets/1/6/CTA_Ashland_full_board_final.pdf
62 "Ashland BRT." Chicago Deparmtnet of Transportation.
https://www.transitchicago.com/assets/1/6/CTA_Ashland_BRT_Fact_sheet_English_FINAL.pdf

delivery vehicles using the lane as a loading space. The CTA's Loop Link, a set of DBLs on Washington, Madison, Clinton, and Canal is indeed no stranger to such trespassing from private vehicles on its lanes[63]. Nevertheless, with better enforcement, DBLs have had success at improving performance and increasing reliability in other bus systems. [64,65, 66]

Currently, only 4 miles of the CTA's 1536 route miles are DBLs. The CTA recently announced plans to add more on Chicago, Western Ave and 79th St as a part of its new Bus Priority Zone Program (BPZ). From the route-level bunching analysis, it is clear that these additions are dearly needed. The 66, 49, and 79 bus routes that run on Chicago, Western, and 79th respectively are ranked second, third, and first in number of bunching incidents over the study period. Based on the results of the route level analysis, eastbound and northbound for each of the routes experienced greater levels of bunching. As such, this study recommends that DBL construction be prioritized for the east and north traveling lanes on the affected streets. All in all, as a smaller investment in new transit infrastructure compared with a comprehensive BRT upgrade, DBLs in these route segments are poised to have significant returns in gains in efficiency for the CTA.

Transit Signal Priority. Of the recommendations made thus far, Transit Signal Priority (TSP) stands out as an inconspicuous, but impactful, method the CTA can implement to reduce bunching incidents. TSP mitigates schedule deviances that arise from dwell time at intersections and traffic lights. TSP works by connecting bus requests for signal priority to the stop-light system at a given intersection[67]. Once the traffic signal receives the request, software integrated with the traffic control system determines how to best respond to the bus's request for

---

[63] Coffey, Chris, and Katy Smyser. "City Not Enforcing Traffic Laws to Help Loop Link Run ..." NBC Chicago, February 17, 2017. https://www.nbcchicago.com/news/local/city-not-enforcing-traffic-laws-to-help-loop-link-run-smoothly-records/31666/.
[64] Abdelfatah, Akmal, and Amro R. Abdulwahid. "Impact of Exclusive Bus Lanes on Traffic Performance in Urban Areas." *Proceedings of the 2nd World Congress on Civil, Structural, and Environmental Engineering*, 2017. https://doi.org/10.11159/icte17.125.

[65] Ben-Dor, Golan, Eran Ben-Elia, and Itzhak Benenson. "Assessing the Impacts of Dedicated Bus Lanes on Urban Traffic Congestion and Modal Split with an Agent-Based Model." *Procedia Computer Science* 130 (2018): 824–29. https://doi.org/10.1016/j.procs.2018.04.071.

[66]Burinskienė, Marija, Modesta Gusarovienė, and Kristina Gabrulevičiūtė-Skebienė. "The Impact of Public Transport Lanes on the Operating Speed of Buses." *The 9th International Conference "Environmental Engineering 2014"*, 2014. https://doi.org/10.3846/enviro.2014.112.

[67] Smith, Harriet, Brendon Hemily, and Miomir Ivanovic. "Transit Signal Priority (TSP): A Planning and Implementation Handbook." *Transportation Research Board,* May 2005.

signal priority. Possible responses include extending a green light or activating a left turn signal early so that the bus can pass through the intersection with minimal stoppage.

In other public transit systems where TSP is widely used, the signaling system has increased schedule adherence and reduced both travel time and delays[68]. TSP is similar to DBLS in that it is typically a component of BRT whose individual implementation presents a smaller cost to budget-conscious public transit systems. In the greater Chicagoland region, an effort to deploy TSP along well-traveled transit roadways has been in progress since 2016. That said, the majority of routes where TSP is planned are suburban Pace bus routes. Thus far, only two corridors in the CTA network, Western and Ashland Ave, are receiving TSP capabilities at signal-controlled intersections. Under the new BPZ program, TSP additions at key intersections on 79th St and Chicago Ave are also planned. Given the consistent number of day-to-day bunching incidents this study identified at intersections on the east side of the city (e.g., North/Clark, Belmont/Clark), there is certainly precedence for the TSP deployment effort to be expanded to transit corridors along the lakefront as well. Another discussion should be held over when buses should receive signal priority. At every level of analysis, the results from the study indicate a temporal bimodality of bus bunching around morning and evening rush hours. There is thus an argument to be made that instead of event-based TSP, where signal priority is given to public transit vehicles only on request from an approaching vehicle, TSP is automatically given to buses at certain time periods throughout the day, starting with the two hours between 7-9 am in the morning and 4-6 pm in the afternoon, which obviously correspond with rush hour periods when both public and private vehicular traffic is high.

A ubiquitous theme underlying all of the recommendations made in this section is the prioritization of public bus transit in Chicago’s larger urban transit ecosystem. Buses are a great compromise between cost and efficacy as a transport mode, but their worthwhileness hinges on whether or not they have the space they need to

[68] Wang, Yinghai, Mark Hallenback, and Jianyang Zheng. “Comprehensive Evaluation of Transit Signal Priority System Impacts Using Field Observed Traffic Data .” University of Washington, June 2008.

operate optimally. With effective implementation, the combination of greater BRT, TSP, and dedicated lanes coverage in the CTA's bus network service area will bring the system one step closer to this goal.

Further Research. The analyses conducted thus far have revealed much about the spatial-temporal patterns of bus bunching in Chicago. Regarding the former, my research thus far focuses on the 'where' and 'when' aspects of bus bunching. Even though regression analysis was a component of the research, the primary objective there was still prediction as opposed to inference. While findings related to where and when bus bunching occurs, and where and when it might occur in the future already allow policymakers to make informed, context-aware, decisions on the regions in Chicago that benefit the most from new public transportation plans, providing information on the 'why' of bus bunching can help add nuance to the recommended policy changes, resulting in data-driven recommendations that are more robust. Furthermore, the CTA's service standards manual outlines some forty-five routes that it considers the network's "key routes." Incorporating those and more into our analysis can not only reinforce what has been found thus far, but also bring to light new insights into the patterns of bus bunching in Chicago over space and time.

Key Routes				
4 Cottage Grove	29 State	54B South Cicero	77 Belmont	95E 93rd-95th
8 Halsted	34 South Michigan	55 Garfield	79 79th	95W West 95th
8A South Halsted	35 31st/35th	60 Blue Island/26th	80 Irving Park	119 Michigan/119th
9 Ashland	47 47th	62 Archer	81 Lawrence	151 Sheridan
12 Roosevelt	49 Western	63 63rd	82 Kimball-Homan	155 Devon
J14 Jeffery Jump	49B North Western	66 Chicago	84 Peterson	
20 Madison	52 Kedzie/California	67 67th-69th-71st	85 Central	
21 Cermak	53 Pulaski	71 71st/South Shore	87 87th	
22 Clark	53A South Pulaski	72 North	90 Harlem	
28 Stony Island	54 Cicero	74 Fullerton	91 Austin	

Table 19. Key Bus Routes on the CTA

References

Abdelfatah, Akmal, and Amro R. Abdulwahid. "Impact of Exclusive Bus Lanes on Traffic Performance in Urban Areas." *Proceedings of the 2nd World Congress on Civil, Structural, and Environmental Engineering,* 2017. https://doi.org/10.11159/icte17.125.

"About Us." CTA. Accessed December 2, 2019. https://www.transitchicago.com/about/

Andres, Matthias, and Rahul Nair. "A Predictive-Control Framework to Address Bus Bunching." *Transportation Research Part B: Methodological* 104 (2017): 123–48. https://doi.org/10.1016/j.trb.2017.06.013.

"Ashland BRT." Chicago Department of Transportation. https://www.transitchicago.com/assets/1/6/CTA_Ashland_BRT_Fact_sheet_English_FINAL.pdf

Ceder, Avishai. *Public Transit Planning and Operation: Theory, Modeling and Practice*. Amsterdam: Elsevier, 2007.

"BACK ON THE BUS: SPEEDING UP CHICAGO'S BUSES." Active Transportation Alliance. Accessed December 10, 2019. http://activetrans.org/sites/files/SpeedingupBuses.pdf.

Bates, John, John Polak, Peter Jones, and Andrew Cook. "The Valuation of Reliability for Personal Travel." *Transportation Research Part E: Logistics and Transportation Review* 37, no. 2-3 (April 2001): 191–229. https://doi.org/10.1016/s1366-5545(00)00011-9.

Beaudoin, Justin, and C.-Y. Cynthia Lin Lawell. "The Effects of Urban Public Transit Investment on Traffic Congestion and Air Quality." *Urban Transport Systems*, 2017. https://doi.org/10.5772/66834.

Ben-Dor, Golan, Eran Ben-Elia, and Itzhak Benenson. "Assessing the Impacts of Dedicated Bus Lanes on Urban Traffic Congestion and Modal Split with an Agent-Based Model." *Procedia Computer Science* 130 (2018): 824–29. https://doi.org/10.1016/j.procs.2018.04.071.

Bharathi, B. Dhivya, B. Anil Kumar, Avinash Achar, and Lelitha Vanajakshi. “Bus Travel Time Prediction: a Log-Normal Auto-Regressive (AR) Modelling Approach.” *Transportmetrica A: Transport Science* 16, no. 3 (January 2020): 807–39. https://doi.org/10.1080/23249935.2020.1720864.

Bowman, Larry A., and Mark A. Turnquist. “Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops.” *Transportation Research Part A: General* 15, no. 6 (1981): 465–71. https://doi.org/10.1016/0191-2607(81)90114-x.

Burinskienė, Marija, Modesta Gusarovienė, and Kristina Gabrulevičiūtė-Skebienė. “The Impact of Public Transport Lanes on the Operating Speed of Buses.” *The 9th International Conference "Environmental Engineering 2014"*, 2014. https://doi.org/10.3846/enviro.2014.112.

Chapman, R. A., and J. F. Michel. “Modelling the Tendency of Buses to Form Pairs.” *Transportation Science* 12, no. 2 (1978): 165–75. https://doi.org/10.1287/trsc.12.2.165

Chatterjee, S., and Ali Hadi. *Regression Analysis by Example*. 5th ed. Hoboken, NJ: Wiley., 2012.

Clewlow, Regina, and Gouri Mishra. “Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States.” *Journal of the Transportation Research Board*, October 2017. https://trid.trb.org/view/1485471

Coffey, Chris, and Katy Smyser. “City Not Enforcing Traffic Laws to Help Loop Link Run ...” NBC Chicago, February 17, 2017. https://www.nbcchicago.com/news/local/city-not-enforcing-traffic-laws-to-help-loop-link-run-smoothly-records/31666/.

Criden , Madelaine . “The Stranded Poor: Recognizing the Importance of Public Transportation for Low-Income Households.” *National Association for State Community Services Programs*, 2008. https://nascsp.org/wp-content/uploads/2018/02/issuebrief-benefitsofruralpublictransportation.pdf

Currie, Graham, and Ian Wallis. “Effective Ways to Grow Urban Bus Markets – a Synthesis of Evidence.” *Journal of Transport Geography* 16, no. 6 (November 2008): 419–29. https://doi.org/10.1016/j.jtrangeo.2008.04.007.

Daganzo, Carlos F. “A Headway-Based Approach to Eliminate Bus Bunching: Systematic Analysis and Comparisons.” *Transportation Research Part B: Methodological* 43, no. 10 (2009): 913–21. https://doi.org/10.1016/j.trb.2009.04.002.

Feng, Wei, and Miguel Figliozzi. “Empirical Findings of Bus Bunching Distributions and Attributes Using Archived AVL/APC Bus Data.” *Icctp 2011*, 2011. https://doi.org/10.1061/41186(421)427.

GUENTHNER, RICHARD P. and KASIMIN HAMAT. “Distribution of Bus Transit On-Time Performance.” Transportation Research Board, 1988. https://trid.trb.org/view/302165

Hall, Randolph, Maged Dessouky, and Quan Lu. “Optimal Holding Times at Transfer Stations.” *Computers & Industrial Engineering* 40, no. 4 (2001): 379–97. https://doi.org/10.1016/s0360-8352(01)00039-0.

Higashide, Steven, and Mary Buchanan. “Who's on Board 2019.” TransitCenter. Accessed December 2, 2019. https://transitcenter.org/publication/whos-on-board-2019/.

Hill, S J, and G A Brunner. “Consumer Conceived Attributes of Transportation.” University of Maryland, 1967.

King, Michael, and Paul Lippens. “Complete Streets Chicago.” *Chicago Department of Transportation, 2013*. https://www.chicago.gov/content/dam/city/depts/cdot/Complete%20Streets/CompleteStreetsGuidelines.pdf

Iliopoulou, Christina, Christina Milioti, Eleni Vlahogianni, Konstantinos Kepaptsoglou, and Javier Sanchez-Medina. “The Bus Bunching Problem: Empirical Findings from Spatial Analytics.” *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018. https://doi.org/10.1109/itsc.2018.8569760.

Jiang, Rui-Sen, Da-Wei Hu, and Xue Wu. “Prediction of Bus Bunching Using Smart Card Data.” *Cictp 2019*, February 2019. https://doi.org/10.1061/9780784482292.122.

Koppisetti, M Venkateswararao, Veeraruna Kavitha, and Urtzi Ayesta. “Bus Schedule for Optimal Bus Bunching and Waiting Times.” *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, 2018. https://doi.org/10.1109/comsnets.2018.8328282.

Lee, Kurt K. T., and Paul Schonfeld. “Optimal Slack Time for Timed Transfers at a Transit Terminal.” *Journal of Advanced Transportation* 25, no. 3 (1991): 281–308. https://doi.org/10.1002/atr.5670250304.

“Night Owl Service.” *Night Owl Service*, 2018. https://www.transitchicago.com/assets/1/6/ctamap_OwlService.pdf.

Perk, Victoria. “Transit Ridership, Reliability, and Retention.” *National Center For Transit Research*, January 2008. https://doi.org/10.5038/cutr-nctr-rr-2005-08.

Pilachowski, Joshua Michael. “An Approach to Reduce Bus Bunching.” University of California, Berkeley, 2009.

Pi, Xidong, Mark Egge, Jackson Whitmore, Amy Silbermann, and Zhen Qian. “Understanding Transit System Performance Using AVL-APC Data: An Analytics Platform with Case Studies for the Pittsburgh Region.” *Journal of Public Transportation* 21, no. 2 (2018): 19–40. https://doi.org/10.5038/2375-0901.21.2.2.

Powell, Warren B., and Yosef Sheffi. “A Probabilistic Model of Bus Route Performance.” *Transportation Science* 17, no. 4 (1983): 376–404. https://doi.org/10.1287/trsc.17.4.376.

Prashker, Joseph N. “Direct Analysis of the Perceived Importance of Attributes of Reliability of Travel Modes in Urban Travel.” *Transportation* 8, no. 4 (1979): 329–46. https://doi.org/10.1007/bf00167987.

Rashidi, Soroush, Prakash Ranjitkar, Orosz Csaba, and Andy Hooper. “Using Automatic Vehicle Location Data to Model and Identify Determinants of Bus Bunching.” *Transportation Research Procedia* 25 (2017): 1444–56. https://doi.org/10.1016/j.trpro.2017.05.170.

Rodríguez, German. “Models for Count Data With Overdispersion” (Nov 2013). https://data.princeton.edu/wws509/notes/c4a.pdf

Saw, Vee-Liem, and Lock Yue Chew. “No-Boarding Buses: Synchronisation for Efficiency.” *Plos One* 15, no. 3 (2020). https://doi.org/10.1371/journal.pone.0230377.

Seneviratne, Prianka N. “Simulation of Fixed Route Bus Travel Time.” *Journal of Advanced Transportation* 22, no. 1 (1988): 39–53. https://doi.org/10.1002/atr.5670220104.

Smith, Gary. “Step Away from Stepwise.” *Journal of Big Data* 5, no. 1 (2018). https://doi.org/10.1186/s40537-018-0143-6.

Smith, Harriet, Brendon Hemily, and Miomir Ivanovic. “Transit Signal Priority (TSP): A Planning and Implementation Handbook.” *Transportation Research Board*, May 2005.

“Top 5 CTA Complaints via CTA Fails - Chicago Tribune.” Chicago Tribune, February 9, 2016. https://www.chicagotribune.com/redeye/redeye-top-5-cta-fails-complaints-20160120-story.html.

“Turnaround: Fixing New York City’s Buses - TransitCenter.” TransitCenter. Accessed December 10, 2019. http://transitcenter.org/wp-content/uploads/2016/07/Turnaround_Fixing-NYCs-Buses-20July2016.pdf.

Wang, Yinghai, Mark Hallenback, and Jianyang Zheng. “Comprehensive Evaluation of Transit Signal Priority System Impacts Using Field Observed Traffic Data .” University of Washington, June 2008. https://www.wsdot.wa.gov/research/reports/fullreports/699.1.pdf

“What Is BRT?” Institute for Transportation and Development Policy. ITDP. Accessed February 9, 2020. https://www.itdp.org/library/standards-and-guides/the-bus-rapid-transit-standard/what-is-brt/.

“Where Workers Work - Labor Market Information.” Illinois Department of Employment Security, November 2019. https://www2.illinois.gov/ides/lmi/pages/where_workers_work.aspx.

“Why Build BRT on Ashland?” Chicago Transit Authority. https://www.transitchicago.com/assets/1/6/CTA_Ashland_full_board_final.pdf

Wisniewski, Mary. “Fewer Chicago-Area Residents Taking Buses, Trains.” chicagotribune.com. Chicago Tribune, May 23, 2019. https://www.chicagotribune.com/news/breaking/ct-biz-transit-ridership-decline-20181130-story.html.

Verbich, David, Ehab Diab, and Ahmed El-Geneidy. “Have They Bunched Yet? An Exploratory Study of the Impacts of Bus Bunching on Dwell and Running Times.” *Public Transport* 8, no. 2 (2016): 225–42. https://doi.org/10.1007/s12469-016-0126-y.

“Your Bus Is On Time. What Does That Even Mean?” TransitCenter, August 13, 2019. https://transitcenter.org/bus-time-even-mean/.

Yu, Haiyang, Zhihai Wu, Dongwei Chen, and Xiaolei Ma. “Probabilistic Prediction of Bus Headway Using Relevance Vector Machine Regression.” *IEEE Transactions on Intelligent Transportation Systems* 18, no. 7 (November 15, 2016): 1772–81. https://doi.org/10.1109/tits.2016.2620483.

Zhao, Jiamin, Maged Dessouky, and Satish Bukkapatnam. “Optimal Slack Time for Schedule-Based Transit Operations.” *Transportation Science* 40, no. 4 (2006): 529–39. https://doi.org/10.1287/trsc.1060.0170.

Zhao, Shuzhi, Chunxiu Lu, Shidong Liang, and Huasheng Liu. “A Self-Adjusting Method to Resist Bus Bunching Based on Boarding Limits.” *Mathematical Problems in Engineering* 2016 (2016): 1–7. https://doi.org/10.1155/2016/8950209.

Appendix

https://drive.google.com/file/d/1WtzUPzmPJGDbcFDV2a1bIDqrl8oqpys_/view?usp=sharing

Fig 7. Time-lapse heatmap of bunching incidents by day

	Dependent variable:				
	bunching_counts				
	OLS	*Poisson*	*Negative binomial*	*Hurdle (count)*	*Hurdle (zero)*
	(1)	(2)	(3)	(4)	(4)
oct12_boardings	0.00004	0.00003	0.0003*	0.0003	0.0003***
	(0.001)	(0.0001)	(0.0002)	(0.0002)	(0.0001)
is_rush	0.009	0.001	-0.018		
	(0.416)	(0.041)	(0.094)		
bg_pop_den	-0.00001	-0.00001***	-0.00001*		
	(0.00001)	(0.00000)	(0.00000)		
bg_percent_transit	0.444	0.181**	0.061		
	(0.937)	(0.087)	(0.209)		
ct_jobdense	-0.00000	-0.00000	0.00000		
	(0.00000)	(0.00000)	(0.00000)		
bg_hhi	0.00001	0.00000***	0.00000***		
	(0.00000)	(0.00000)	(0.00000)		
DIRNB	-0.306	-0.156***	-0.380***	-0.393***	-0.030
	(0.386)	(0.031)	(0.083)	(0.145)	(0.094)
DIRSB	-0.755*	-0.412***	-0.517***	-0.471***	-0.358***
	(0.389)	(0.036)	(0.085)	(0.140)	(0.096)

DIRWB	-1.017***	-0.723***	-0.875***	-1.223***	-0.294***
	(0.390)	(0.039)	(0.087)	(0.142)	(0.096)
POSFT	0.253	0.186*	-0.006	0.776*	-0.493*
	(1.040)	(0.101)	(0.241)	(0.455)	(0.278)
POSMB	-0.053	-0.053	-0.208	0.091	-0.270*
	(0.649)	(0.065)	(0.151)	(0.245)	(0.163)
POSMT	2.589***	1.251***	0.776***	2.038***	-0.673**
	(0.968)	(0.062)	(0.208)	(0.462)	(0.265)
POSNS	0.332	0.227***	0.366***	0.450***	0.077
	(0.304)	(0.029)	(0.068)	(0.117)	(0.075)
POSNT	-0.150	-0.144*	-0.082	0.308	-0.359**
	(0.680)	(0.077)	(0.159)	(0.280)	(0.175)
POSTERM	1.658	0.975***	1.115***	1.084***	0.714***
	(1.113)	(0.081)	(0.232)	(0.362)	(0.265)
rts_served	0.158	0.072***	0.036		
	(0.124)	(0.010)	(0.027)		
lag_term	1.678***	1.128***	1.201***	0.671***	1.585***
	(0.276)	(0.027)	(0.061)	(0.105)	(0.067)
oct12_boardings:is_rush	0.0001	0.0001	0.00001		
	(0.001)	(0.0001)	(0.0002)		
is_rush:bg_pop_den	0.00001	0.00001***	0.00001		
	(0.00002)	(0.00000)	(0.00000)		
Constant	0.443	-0.426***	-0.343***	-15.492	-0.895***
	(0.560)	(0.055)	(0.125)	(303.073)	(0.095)
Observations	4,380	4,380	4,380	4,380	
R^2	0.015				
Adjusted R^2	0.010				
Pseudo R^2		.13	.15	.10	

Log Likelihood		-13,355.750	-6,287.994	-6,020.400
theta			0.347^{***} (0.012)	
Akaike Inf. Crit.		26,751.500	12,615.990	12,090.8
Residual Std. Error	8.835 (df = 4360)			
F Statistic	3.406^{***} (df = 19; 4360)			

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 13. Model Comparison